INTERNATIONAL REVIEW OF CYTOLOGY, VOL. 76

Highly Repeated Sequences in Mammalian Genomes

MAXINE F. SINGER

Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

I. Introduction 67
II. Perspectives on Methods 68
III. Satellites 71
A. Changing Concepts 71
B. Rodents 73
C. Bovine 77
D. Primates 80
E. Marsupials 86
F. Carnivora 86
G. Comparisons between Satellite Sequences 86
H. Chromosomal Location 87
I. The Question of Satellite Function 88
J. Amplification of Satellites 89
IV. Interspersed Repeated Sequences 91
A. Emerging Concepts 91
B. Short Interspersed Repeated Sequences (SINES) 92
C. Long Interspersed Repeated Sequences (LINES) 102
D. Amplification and Dispersion 104
V. Concluding Remarks 105
References 106
Note Added in Proof 112

I. Introduction

The discovery of restriction endonucleases and the development of molecular cloning and DNA sequencing techniques have revolutionized the study of the structure of complex genomes. The unexpected insights generated by these new techniques have created profound changes in our understanding of the way genetic information is organized and expressed. Consequently, many new concepts regarding the highly repeated DNA sequences in eukaryotic genomes are rapidly developing. Perhaps the only generalization that can safely be made at this time is that both the repeated sequences and their genomic organization are much more complex than previously suspected. This article attempts to summarize the emerging picture regarding the structure and organization of mammalian highly repeated sequences at the molecular level. Both the tandemly repeated sequences

ISBN 0-12-364476-3

widely termed "satellites" and those segments that are interspersed among other genomic DNA sequences will be described. As a general rule of thumb, any sequence repeated more than 10^4 times is included. I have chosen to concentrate on mammals because of my own interests and because other recent reviews emphasize plants (Bedbrook and Gerlach, 1980; Flavell, 1980) and invertebrates including crabs (Christie and Skinner, 1980), *Drosophila* (Appels and Peacock, 1978; John and Miklos, 1979; Brutlag, 1980; Hilliker *et al.*, 1980; Spradling and Rubin, 1981), and sea urchins (Moore *et al.*, 1980; Posakony *et al.*, 1981). The available information is for the most part structural. Unfortunately, the function, if any, of most of these sequences remains an enigma.

Most articles covered here were published prior to the end of July 1981 although I had, through the courtesy of colleagues, preprints of other papers. My own interests dictated selection of the material. The interesting matter of the organization of highly repeated sequences within chromatin is discussed extensively in several recent reviews (Zachau and Igo-Kemenes, 1981; Kornberg, 1981; Igo-Kemenes *et al.*, 1982) and is not included here. Repeated sequences that are known to be genes (e.g., histone and ribosomal RNA genes) have also been reviewed recently (Long and Dawid, 1980). There are available several summaries of current thinking about the evolution of highly repeated sequences (Brutlag, 1980; Bostock, 1980; Dover, 1981). My own bias is that while relevant structural information is rapidly accumulating, we are still largely ignorant of overriding principles. In particular, with the question of function unresolved, discussions of evolution are perforce carried out in the absence of any sense of selective pressure, a critical defect.

Throughout this article the words "repeated" and "reiterated" are used interchangeably. Neither one implies that the many copies of a given family of sequences are identical. Unless demonstrated otherwise, all "repeats" are assumed to be members of a set of closely similar but somewhat variant DNA segments; "closely similar" means that they hybridize to one another under stringent conditions (0.45 *M* NaCl, greater than 65°C). The term "consensus sequence" will be used to describe a nucleotide sequence representing the most abundant nucleotide at each position in the repeat units comprising a set.

II. Perspectives on Methods

In the past, two methods dominated analysis of repeated DNA sequences: measurement of DNA renaturation kinetics (Britten *et al.*, 1974) and isopycnic centrifugation in gradients of CsCl and $CsSO_4$ (Szybalski, 1968). Gross physical separation of highly repeated, middle repeated, and unique sequences was achieved by taking advantage of (1) the dependence of renaturation rate on the concentration of a sequence, (2) the ability of hydroxyapatite to separate single-

and double-stranded DNA, and (3) the separation of some satellite sequences from bulk DNA by virtue of a unique density. Together, these methods demonstrated that (1) eukaryote genomes can be divided up into classes of DNA sequences according to the reiteration frequency (Britten and Kohne, 1968), (2) many very highly repeated sequences are in long tandem arrays (satellites), and (3) some repetitive sequences are dispersed throughout major portions of genomes amid either other repeated sequences or sequences present only once per genome, that is, "unique" sequences.

Even before molecular cloning came into wide use, it was apparent that these useful characterizations masked much greater complexity. For example, the mixture of sequences in the "middle repeated" category includes segments repeated anywhere from two times to tens or hundreds of thousands of times. Some of these sequences proved to be functional genes—either identical multicopy genes such as those for ribosomal RNA and histones, or closely related genes encoding similar but distinctive gene products such as the families of actin, β-globin, or immunoglobulin genes (see review by Long and Dawid, 1980). Furthermore, because divergence among the members of a set of repeated sequences decreases the rate at which they reassociate, middle repeated sequences can appear in the single copy class and the copy number of highly repeated sequences is easily underestimated. Similarly, although many organisms show one or more discrete satellite DNA fractions apart from the main band upon isopycnic centrifugation, others yield no such fractions. Centrifugation in the presence of heavy metals or various antibiotics or dyes, tease "cryptic satellites" out of the DNA of many organisms, but others yield no satellite even though renaturation kinetics shows the presence of highly repeated sequences.

New methods have now begun to unravel some of the complexities and it is the application of these methods that is the basis for the observations summarized in this article. Restriction endonucleases cleave DNA molecules after recognition of specific short nucleotide sequences. The recognition and cleavage sites of some of the enzymes referred to in this article are shown in Table I. Complete current lists of known restriction endonucleases are available (Roberts, 1980). Digestion of DNA with a specific enzyme reproducibly divides it up into a set of fragments whose sizes depend on the spacing between recognition sites. The mixture of fragments can be conveniently separated according to size by electrophoresis on semisolid supports such as agarose or polyacrylamide (Southern, 1979). When the gels are stained to mark DNA the mixture appears as a continuous smear of fragments of all possible sizes. However sequences that are repeated sufficiently to represent at least 0.5% of the genome often appear as distinct bands against the background smear. The entire collection of fragments on the gel can be transferred without disturbing their distribution by blotting after denaturation onto sheets of nitrocellulose (Southern, 1975b, 1979) or diazotized paper (Alwine *et al.*, 1979). Thereafter, the sheets can be incubated in the presence of radioactive

TABLE I
RECOGNITION AND CLEAVAGE SITES OF SOME RESTRICTION ENDONUCLEASES[a]

Enzyme	Sites
*Sau*96I	5′-G′G N C C - -C C N G′G -5′
*Ava*II	A 5′-G′G (T)C C - -C C (T)G′G -5′ A
*Taq*I	5′-T′ C G A - -A G C′ T -5′
*Sau*3A *Mbo*I	5′-′G A T C - -C T A G′-5′
*Eco*RI	5′-G′A A T T C- -C T T A A′G-5′
*Alu*I	5′-A G′C T - -T C′G A -5′
*Hind*III	5′-A′A G C T T- -T T C G A′A-5′
*Bam*HI	5′-G′G A T C C- -C C T A G′G-5′

[a] Taken from Roberts (1980) where a complete list can be found. The indicates the site of cleavage within the recognition site. Enzymes with identical specificity (*Sau*3A and *Mbo*I) are called isoschizomers.

DNA or RNA fragments to permit hybridization of the probe with homologous sequences on the sheet. Exposure of the sheet to X-ray film provides an autoradiogram on which a darkened band reveals the DNA fragments homologous to the probe. These techniques are sufficiently sensitive to reveal even unique DNA sequences amid fragments generated from an entire mammalian genome.

Molecular cloning, usually in *E. coli* host vector systems, provides ways to purify and amplify DNA segments generated after restriction endonuclease digestion or after random generation of genomic fragments by shearing or nonspecific nuclease digestion (Morrow, 1979). Hybridization techniques with radioactive probes permit identification of the clones of interest and primary nucleotide sequence is readily determined (Maxam and Gilbert, 1980; A. J. H. Smith, 1980). In this way, single molecular species corresponding to the individual members of a repeated set are available as are segments corresponding to a tandem array or a dispersed repeated sequence enclosed within its natural neighbors. One very important consequence of these methods is the availability of pure DNA segments for use as probes in the analysis of genome organization by the blotting procedures described above. Uncloned probes are likely to be

contaminated by extraneous DNA sequences. Because of the very high sensitivity of the blotting methods even minor contaminants lead to misleading data. A word of caution: molecular cloning is not without problems. Many investigators have noted that deletions often occur when tandemly repeated arrays are cloned and amplified in *E. coli* (e.g., Brutlag *et al.*, 1977; Carlson and Brutlag, 1977; Sadler *et al.*, 1980; Sakano *et al.*, 1980).

III. Satellites

A. Changing Concepts

The characteristic organizational feature of satellites and cryptic satellites is the tandem repetition of a unit DNA sequence. In accordance with a previous suggestion (Pech *et al.*, 1979b), the term satellite will be used to describe such DNA regions regardless of whether or not they are separable as classical satellites by isopycnic centrifugation. This usage, while not completely accurate, conforms with widespread practice. Satellites comprise anywhere from a few percent (human) to over 50% (Kangaroo rat) of mammalian genomes. Common characteristics of satellites include (1) association with heterochromatin, (2) lack of measurable transcription (but see below), (3) replication late in S-phase, and (4) underreplication in polytene chromosomes (see Brutlag, 1980, for review).

At one time it was believed that all satellites contain simple redundant repeats of short oligonucleotide segments. However molecular analysis has demonstrated that repeat units vary from a few to several thousand base pairs and that enormous complexity resides within these sequences. Satellite arrays frequently resist separation by isopycnic centrifugation and remain instead within the main density fraction of genomic DNA. This is true even when some satellite is separable by centrifugation; additional satellite may remain sequestered in the main band. Furthermore, neither the purity nor the unique character of an isopycnic satellite fraction can be assumed. Sequences present in one satellite band may also occur in others and in main band. Also several nonhomologous repeat units may reside in one density satellite fraction, either linked in one molecule or on separate molecules.

Regardless of whether separated satellite or total DNA is used, both the repeat unit and its tandem organization can be revealed by restriction endonuclease digestion. When a site for a particular restriction endonuclease occurs within a typical repeat unit, digestion converts a tandem array to a set of DNA fragments of repeat unit length (a type "A" digestion; Hörz and Zachau, 1977). Partial digests generate "ladders" of fragments that are demonstrable upon electrophoresis. The "ladder" fragments are integral multiples of the basic repeat unit in length, and provide evidence for the tandem organization of the satellite.

After exhaustive digestion, a "ladder" of resistant multiples usually remains. The resistant ladder arises because of sequence alteration at the canonical restriction endonuclease site in occasional copies of the repeat unit. Such alterations may be (1) randomly dispersed among the copies of the unit repeat, or (2) clustered within neighboring repeats, or (3) occur at regular intervals along an array. All three possibilities occur. The latter two nonrandom arrangements (and combinations of the two) define distinctive subarrays of a satellite and these are here called satellite domains. Restriction endonuclease sites that are missing from the typical unit repeat sequence may occur occasionally within an array because of sequence variation and yield distinctive digestion patterns with a preponderance of very long fragments (a type "B" digestion). Again the same 3 distributions occur and may define specific satellite domains. The absence of an A-type site or the presence of a B-type site at a fixed frequency within an array of repeats is seen as a long repeat length superimposed on a shorter canonical reiterated unit.

The primary nucleotide sequence of a repeat unit can be determined using an entire set of monomeric units generated by type A digestion or a molecularly cloned member of the set. Unambiguous consensus sequences are frequently obtained with uncloned sets, indicating that variant bases occur in fewer than 15% of the copies at any given position. The 15% limit reflects the limits of detectability in the DNA sequencing procedures. In the typical case one restriction endonuclease is used to generate monomer units; any repeat units lacking the site appear in the higher ladder and are not represented in the consensus. Subsequent steps in the preparation of fragments for sequencing often utilize additional restriction sites in the repeat unit to generate subunit size sections. At each such cleavage repeat units lacking the site in question will be discarded with undigested material and will not be represented in the consensus. Thus, the existence of subsets of a given monomer bias the sequence data on uncloned fragments.

Some organisms have multiple distinguishable satellites. Sequence analysis frequently shows that these may be related to one another in spite of different densities and restriction endonuclease cleavage patterns or inability to cross hybridize. Other organisms appear to have one predominant satellite. In these cases, molecular analysis may show the presence of domains. It is possible that domains and different satellites are organizationally equivalent, representing localized amplifications of particular variants of a basic sequence. There is growing evidence that different domains or distinguishable satellites may tend to be localized to specific chromosomes. Both the domains of a relatively homogeneous satellite and the several satellites in some organisms can all be seen as the product of alternating cycles of mutation (including single base pair changes, deletions, and insertions) and amplification and deletion. This scheme,

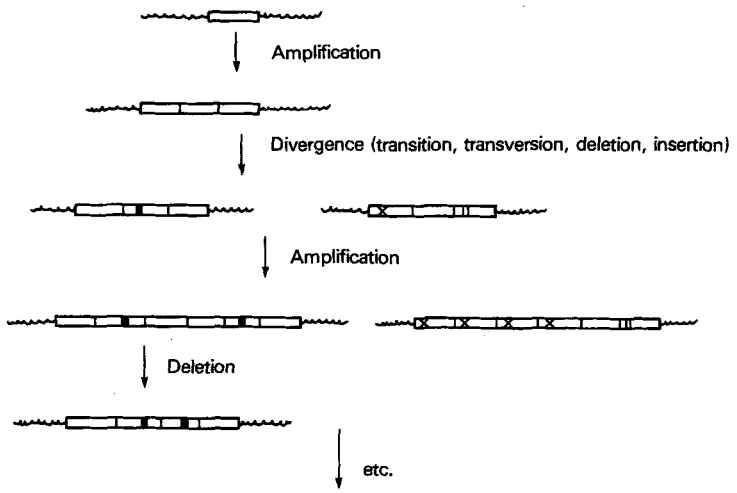

FIG. 1. Schematic diagram describing the generation of related but divergent satellites, including domains, by alternating cycles of mutation and amplification (after Southern, 1970).

first suggested by Southern (1970), is outlined in Fig. 1; it is consistent with the data on various satellites, as reviewed below.

B. RODENTS

1. *Mouse*

The single density satellite of *Mus musculus* makes up 5 to 10% of the genome (Kit, 1961) and is localized at centromeric heterochromatin in all mouse chromosomes except the Y chromosome (Pardue and Gall, 1970). Most (60–70%) of isolated mouse satellite is digested to a set of 234-bp-long monomeric units by *Sau*96I (Hörz and Altenburger, 1981), *Eco*RII (Southern, 1975a; Hörz and Zachau, 1977), and *Ava*II (Dover, 1978). The set yields an unambiguous consensus sequence (Hörz and Altenburger, 1981; Manuelidis, 1981a) (Fig. 2a). The asymmetric distribution of A and T residues on the two DNA strands of the repeat unit may account for anomalous prior estimates of both the size of the repeat unit (by gel electrophoresis) and of the GC content (by density gradient centrifugation or melting temperature determination). The 234-bp segment comprises four related internal tandem repeats—58, 60, 58 and 58 bp in length, respectively. And each of these can be further divided into two related but variant 28- and 30-bp-long segments. Thus the 234-bp repeat encloses 8 shorter tandem repeats. Hörz and Altenburger (1981) deduced a common progenitor sequence

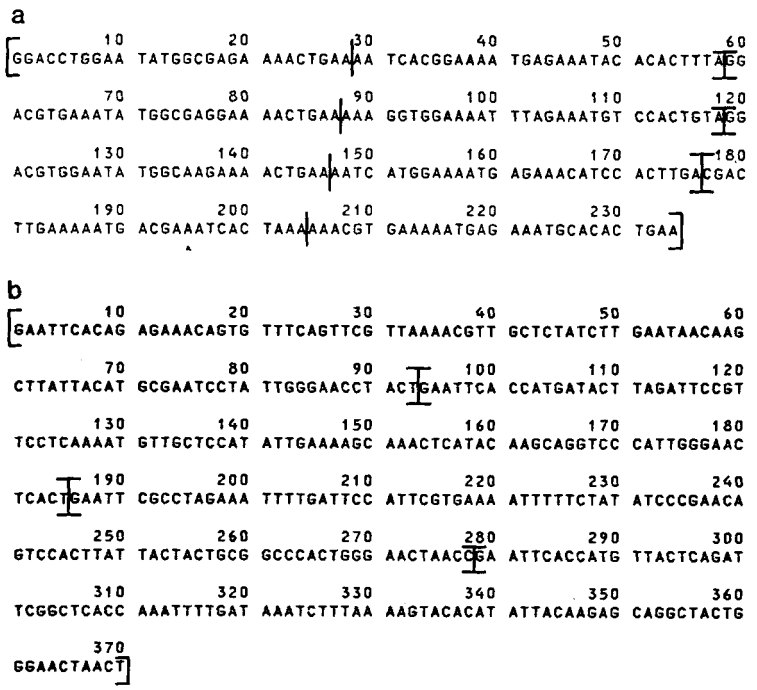

FIG. 2. The consensus sequences of the major monomeric repeat units in (a) mouse satellite (Hörz and Altenburger, 1981) and (b) rat satellite I (Pech *et al.*, 1979b). Brackets enclose the long internal repeats and lines mark off the 8 shorter internal repeats in the mouse sequence. Only one DNA strand is shown and it is 5′ to 3′ reading from left to right.

from the structure of the 8 repeats and suggested that the satellite may have been constructed by alternating mutation and amplification starting with three similar nonanucleotides: GAAAAATGA, GAAAAAACT, GAAAAACGT.

Several variants of the 234-bp consensus sequence have been identified. Thus, approximately 5–10% of the 234-bp-long fragment set contains a single *Taq*I cleavage site that is missing in the consensus sequence (Hörz and Altenburger, 1981). Approximately 20% of the satellite is degraded to dimers (464 bp) after exhaustive cleavage with either *Sau*96I or *Eco*RII (Hörz and Zachau, 1977; Southern, 1975a) and thus variation at these (overlapping) restriction sites (residues 1–8 in Fig. 2a) is frequent; the dimer yields a consensus sequence almost identical to that of the monomer (Hörz and Altenburger, 1981). *Sau*96I digestion also yields small amounts of heretical fragments $0.25 \times n \times 234$ bp in length (n is an integer); these structures have not been studied in detail. There is extensive methylation of C-G sequences in the satellite (Hörz and Altenburger, 1981). Further, some sequences that hybridize with mouse satellite are found in DNA

from which satellite has been removed by density gradient centrifugation (Manuelidis, 1980; Stambrook, 1981).

Individual mouse chromosomes may carry distinct satellite domains. DNA from a Chinese hamster/mouse hybrid cell line containing only the mouse X chromosome in 75% of the cells was studied by restriction endonuclease digestion and hybridization with ^{32}P-labeled density gradient purified satellite (Brown and Dover, 1980a). There is no cross-reaction with the Chinese hamster DNA. The data indicate that the X-chromosome contains mouse satellite domains whose basic organization is similar to the bulk of the satellite but that (1) contain fewer *Hin*f, *Alu*I, and *Eco*RI sites, and (2) lack heretical size fragments.

Sequences homologous to the *M. musculus* satellite have been detected in other *Mus* species by hybridization with radioactive probes of *M. musculus* satellite purified by density gradient centrifugation; the data require confirmation with cloned probes. There was less *in situ* hybridization to *M. booduga* chromosomes than to those of *M. musculus* although the distribution was similar; only one segment on the short arm of the X-chromosome in *M. dunni* showed significant hybridization (Sen and Sharma, 1980). Hybridization to restriction endonuclease fragments of total *M. spretus* DNA (Brown and Dover, 1980b) indicated sequences homologous to *M. musculus* satellite at about 1% the level even though *M. spretus* does not yield a classical density satellite. Almost all the homologous material in *M. spretus* was degraded by *Ava*II, as is *M. musculus* satellite. *Alu*I, which degrades about 10% (type B) of *M. musculus* satellite, digested only about 2–3% of the homologous sequences in *M. spretus* to separable products suggesting different subsets of sequences in the two species. Restriction endonuclease digests suggest the presence of related satellite sequences in two species of the field mouse genus *Apodemus* (Brown and Dover, 1979).

2. *Rat*

No substantial satellite, cryptic or otherwise, can be isolated from total *R. rattus* DNA by isopycnic centrifugation, although renaturation kinetics shows that almost 10% of the rat genome comprises highly repeated sequences (Bonner *et al.*, 1973). Digestion of total rat or Novikoff hepatoma ascites cell DNA with several restriction endonucleases yields ladders of DNA fragments of defined chain length, indicative of the presence of tandem repeated sequences (Philippsen *et al.*, 1974; Maio *et al.*, 1977; Fuke and Busch, 1979; Lapeyre and Becker, 1980; Sealy *et al.*, 1981). Some distinct differences in the pattern of *Eco*RI bands observed with rat liver nuclear and Novikoff hepatoma DNAs have been reported (Lapeyre and Becker, 1980). A predominant 370-bp-long *Hin*dIII band is degraded by *Eco*RI to yield bands about 93 bp long, which is the size of the smallest abundant band obtained by *Eco*RI digestion of total rat DNA. About 3% of the rat genome is accounted for by these fragments (Fuke *et al.*, 1979;

Pech *et al.*, 1979b) and the sequences may be concentrated in nucleolar DNA (Fuke *et al.*, 1979).

Approximately 1–3% of rat DNA is converted to fragments about 10 kbp in length by digestion with *Sau*3A (type "B" digestion) while most of the rest of the genome is simultaneously degraded to a large mixture of small fragments whose average chain length is 400 bp (Pech *et al.*, 1979b). The 10-kbp fraction was purified by preparative gel electrophoresis and named rat satellite I. Analysis of the purified satellite I (Pech *et al.*, 1979b) demonstrated a repeat length of 370 bp constructed of four related internal segments 93, 92, 93, and 92 bp in length, respectively (Fig. 2b). Each of the internal segments contains an *Eco*RI site (at residues 1–6, 94–99, 186–191, and 279–284 in Fig. 2b) and one contains a *Hin*dIII site (residues 58–63). The sequence data obtained with the set of *Hin*dIII monomers (Fig. 2b) revealed frequent variations at eight defined residues. These must represent alterations in at least 5–10% of the members of the set. Thus a substantial number of the members diverge in specific nonrandom ways. It is not known whether the different abundant variants are present in the same or different copies of the sequence. Domains have not been characterized, but they may occur since some portion of the sequence is left uncut by each of the restriction endonucleases with otherwise regular sites. The CpG sequences in the satellite are extensively methylated (Pech *et al.*, 1979b). Independent base sequence determination on *Eco*RI fragment sets 92 and 93 bp in length, respectively, and generated by cleavage of total rat DNA with *Eco*RI confirms the relatedness of the *Eco*RI fragments to rat satellite I (Lapeyre *et al.*, 1980; Sealy *et al.*, 1981).

As with *R. rattus* neither satellites nor cryptic satellites can be isolated from *R. norvegicus*, *R. sordidus*, or *R. villosissimus* by isopycnic centrifugation. Also, less than 5% of the three genomes are highly repeated segments (Miklos *et al.*, 1980). Analysis by restriction endonuclease digestion is consistent with the presence of very low amounts of satellite DNA in these three species.

3. *Other Rodents*

The oligomer 5′-TTAGGG-3′ is repeated frequently in a major satellite, Hs, of *Dipodomys ordii*, the kangaroo rat (Fry and Salser, 1977) and in guinea pig α-satellite (or satellite I) (Southern, 1970). The same oligomer probably occurs frequently in *Thomomys bottae*, the pocket gopher, *Ammospermophilus leucurus*, the antelope ground squirrel, and other species in the genus *Dipodomys* (Fry and Salser, 1977; Mazrimas and Hatch, 1977). Analysis of the total repetitive DNA of the guinea pig, *Cavia porcellus*, was made by isolating the fraction reassociating at a C_0t of about 7×10^{-2} and removing remaining single strands with S1 nuclease (Hubbell *et al.*, 1979). The data suggested that about 21% of the genome of *C. porcellus* is highly repeated while earlier estimates, based on density gradient isolation of satellites indicated about 10%. It is possible that the

rapidly annealing fraction contains both satellites and interspersed repeated sequences as it is heterogeneous. The three *C. porcellus* satellites that are separable by isopycnic centrifugation were used to prepare ^{3}H-labeled cRNAs and the probes were hybridized *in situ,* to *C. porcellus* chromosomes (Duhamel-Maestracci *et al.,* 1979). All hybridization was to centromeric regions and some distributional specificity of the 3 satellites was observed; the Y chromosome hybridized to none of the three probes. Confirmation of the data with cloned probes is required.

C. Bovine

The eight different satellites that have been distinguished in calf thymus DNA (Macaya *et al.,* 1978; Kopecka *et al.,* 1978) are listed in Fig. 3 along with their buoyant densities in CsCl, their relative abundance in the genome, and, where known, a schematic diagram of the organization of the repeat units. The purification of these satellites depends on sophisticated use of differential density gradient centrifugation in Cs_2SO_4 in the presence of various additives (Macaya *et al.,* 1978; Streeck *et al.,* 1979). Together, the eight comprise over 23% of the genome although analysis of the kinetics of renaturation of calf thymus DNA suggested that less than 5% was in very rapidly renaturing components. This dramatic example emphasizes that satellite and "rapidly renaturing" are not synonymous. Recent analysis of bovine satellites on the molecular level illus-

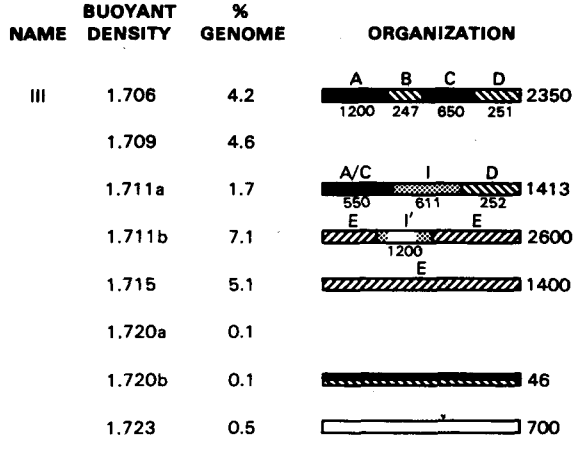

FIG. 3. Bovine satellites and their structural organization where known. See the text for references. The earlier "names" of the bovine satellites are given in the first column; the identity of II is uncertain. Closely related segments repeated in more than one satellite are indicated graphically and by letter. The numbers to the right are the sizes, in base pairs, of the overall repeating units. The sizes of different segments are indicated below. The drawings are not to scale.

trates how distinctions based on satellite density or even restriction endonuclease digestion can obscure striking similarities. Also, these studies show that the size of repeat units determined by restriction endonuclease digestion is not a reliable indicator of underlying structure. In one case, the 1.706 gm/cm^3 bovine satellite, what initially seemed a complex satellite with a repeat unit of 2350 bp, has turned out to have true repeat unit about 0.01 times that size! And at least 5 of the 8 satellites are related in part to one another.

Digestion of purified 1.720b satellite with several different restriction endonucleases yields a series of fragment sets that are multiples of 46 bp in length [Streeck and Zachau, 1978 (note that the 1.720b satellite is erroneously termed 1.723 in that paper), Pöschl and Streeck, 1980]. The set of 46-bp fragments produced by *Alu*I gave an unambiguous consensus sequence (Fig. 4) and revealed an underlying periodicity of two related 23-bp repeats. Individual copies of the repeat units diverge from the consensus sequence. Whether or not particular divergent family members are collected in domains is not clear. The sequence contains several CG dinucleotides and these are probably frequently methylated.

Although quite distinct by density and restriction endonuclease patterns, the

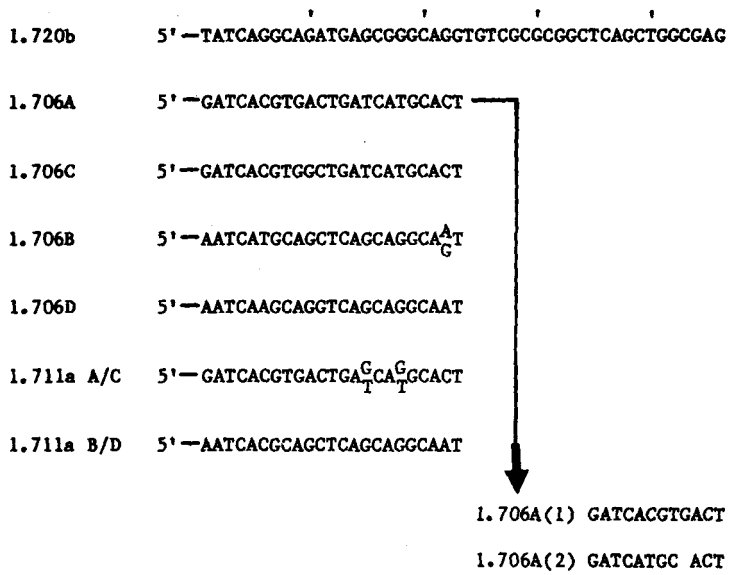

FIG. 4. Sequence relations among the bovine satellites (see Fig. 3). The 46-bp consensus unit of the 1.720b satellite is from Pöschl and Streeck (1980); the 23-bp consensus units of 1.706A, B, C, and D are from Pech *et al.* (1979a); the 23-bp consensus units of 1.711a A/C and B/D are from Streeck (1981). Only one DNA strand is shown in each case. The data are somewhat simplified and the reader is referred to the original papers for details.

sequence of the 1.706 satellite (III) is related to that of the 1.720b satellite. The overall repeating unit of the 1.706 satellite (Streeck and Zachau, 1978; Streeck *et al.*, 1979; Pech *et al.*, 1979a) contains four distinct regions totaling 2350 bp. The distribution of restriction endonuclease sites and direct nucleotide sequencing of both cloned and uncloned fragments revealed that the four regions comprise two pairs of closely related alternating segments (A and C, B and D). The two kinds of segments (A/C, B/D) are themselves distantly related. The underlying unity is the tandem repetition of a 23-bp unit that is related to the repeat unit of the 1.720b satellite (Fig. 4). Furthermore, the 23-bp repeat unit of 1.706 A, for example, itself looks like a tandem repeat of a unit half its size (see Fig. 4).

The 1.711a satellite (Streeck, 1981) is also related to 1.720b and 1.706; an unambiguous consensus sequence was obtained with an uncloned fragment set. The long-range repeat length of 1.711a is 1413 bp and can be divided into 3 regions. Two are marked A/C and D in Fig. 3 to reflect their marked similarity to the corresponding segments in the 1.706 satellite, and the third region, which is unrelated to A/C or D and is not internally repetitive, is marked I. The I region has several features reminiscent of transposable elements (reviewed by Calos and Miller, 1980) including terminal direct and inverted repeats (Streeck, 1981). Indeed, the 1.711a repeat unit can be viewed as one portion of the 1.706 unit interrupted by the I region. Furthermore, I contains open reading frames suitable for translation into protein and a segment similar to the TATAAATA box that is one component of eukaryotic promotors.

More than 95% of the 1.715 satellite from either calf thymus or bovine kidney cells in tissue culture is degraded by *Eco*RI and by *Sal*I to a fragment set about 1400 base pairs in length (Botchan, 1974; Roizes, 1974, 1976; Philippsen *et al.*, 1974; Lipchitz and Axel, 1976; Gaillard *et al.*, 1981). Analysis of the 1400-bp monomer set with other restriction endonucleases established that the bulk of the set shares similar sequences but that variant forms exist (Roizes, 1976; Roizes *et al.*, 1980). At least four distinguishable domains may occur. Superimposed on this nonrandom division into domains other variations appear to reflect random base pair changes leading fortuitously to altered (gain or loss) restriction sites (Roizes *et al.*, 1980). The sequence of the 1.715 satellite monomer does not contain tandemly reiterated internal repeats like the 23 bp regularity found in satellites 1.720 and 1.711a but does contain multiple repeats of short segments of the consensus sequences of those satellites (Roizes *et al.*, 1980; Gaillard *et al.*, 1981; A. Plucienniczak, J. Skowronski, and J. Jaworski, personal communication). Comparison of restriction sites in the cloned and sequenced monomer with those in the bulk satellite suggests extensive methylation (Roizes *et al.*, 1980; Kaput and Sneider, 1979; Gaillard *et al.*, 1981). The extent of methylation appears to differ markedly in sperm and thymus DNA (Sturm and Taylor, 1981).

The 1.711b satellite is similar to 1.715 except that the 1400-bp repeat unit is interrupted about 60% of the time by a 1200-bp-long insertion (I′ on Fig. 3). The arrangement of 1.711b repeat units with and without the I′ element has no known pattern. I′ shares sequence homologies with the I element in 1.711a (Streeck, 1981). The I′ region of 1.711b has inverted terminal repeated sequences and is flanked by duplications of the 6 base pairs at the point of interruption of the 1400-bp repeats. In contrast, the I region in 1.711a is not surrounded by direct repeats of the target sequence.

D. Primates

1. *α-Satellites*

At least one satellite in every primate that has been investigated is part of a kinship of sequences referred to as α-satellite (or alphoid) after the prototypical African green monkey (AGM) *(Cercopithecus aethiops)* satellite (Maio, 1971; Kurnit and Maio, 1974). The AGM α-satellite represents about 20% of the genome (Maio, 1971; Fittler, 1977; Singer, 1979), hybridizes to centromeres (Kurnit and Maio, 1973, 1974; Segal *et al.*, 1976), and comprises a substantial amount of nucleolar DNA (Kurnit and Maio, 1973). Both *in situ* hybridization to chromosomes (Segal *et al.*, 1976) and kinetic analysis of interspersion frequency (Singer, 1979) suggested that some copies of the satellite might be dispersed into other parts of the genome. However this suggestion has not been confirmed on a molecular basis and the kinetic data may reflect the interspersion of *Alu* sequences (see below) rather than α-satellite. Restriction endonuclease analysis and hybridization experiments suggest that related but nonidentical α-satellites are present in *Gorilla gorilla, Pan troglodytes, Cercocebus aterrimus, Macaca mulatta, Mandrillus sphinx, Papio cynocephalus, Cercopithecus pygerythrus, Colobus badius,* and *Cebus capuchinus* (Donehower and Gillespie, 1979; Singer and Donehower, 1979; Gillespie, 1977; Musich *et al.*, 1980; Maio *et al.*, 1981a). Typically, digestion of total or satellite DNA with one or another restriction endonuclease yields a ladder of fragments that are integral multiples of approximately 170 bp in length.

Sequence analysis confirms the presence of related α-satellites in AGM (Rosenberg *et al.*, 1978), baboon (*Papio papio*) (Donehower *et al.*, 1980), bonnet monkey (*Macaca radiata*) (Rubin *et al.*, 1980a), and human (*Homo sapiens*) (Manuelidis and Wu, 1978; Wu and Manuelidis, 1980) DNA (Fig. 5). The length of the basic repeat unit is 172 bp in the AGM and close to twice that in the three other species. The $2n$ units are composed of two different variations of the basic repeat unit (an a-b-a-b-a-b type structure, see legend to Fig. 5). No regular internal periodicity has been discerned although large portions of the sequences are alternating blocks of purines and pyrimidines (Rosenberg *et al.*, 1978) and

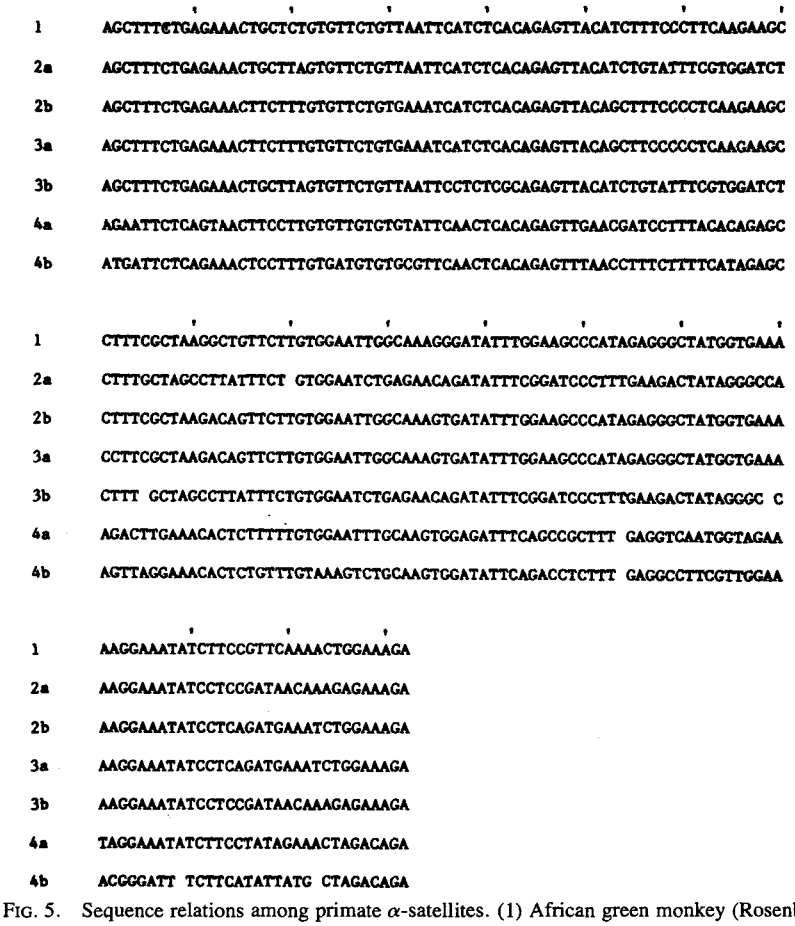

FIG. 5. Sequence relations among primate α-satellites. (1) African green monkey (Rosenberg *et al.*, 1978); consensus sequence of uncloned *Hin*dIII monomer (172 bp). The sequence is displayed from one *Hin*dIII site to the next. (2) Baboon (Donehower *et al.*, 1980); consensus sequence of the dimeric uncloned *Bam*HI monomer (343 bp). The two portions should be read a then b. The sequences are aligned with the African green monkey sequence. The *Bam*HI cleavage site is between residues 115/116 in a. (3) Bonnet monkey (Rubin *et al.*, 1980); consensus sequence of the dimeric uncloned *Hae*III monomer (343 bp). The two portions should be read a then b. The *Hae*III cleavage site is between residues 137–138 in b. (4) Human (Wu and Manuelidis, 1980); consensus sequence of the dimeric uncloned *Eco*RI monomer (340 bp). The two segments should be read a then b. The *Eco*RI cleavage site is between residues 2/3 in a. In each case only a single strand is shown; it reads 5′ to 3′, left to right.

there is a high frequency of 5′-GAAA and 5′-CTTT on the displayed strands (Donehower *et al.*, 1980; Rubin *et al.*, 1980a).

The sequences of cloned monomers of the AGM satellite indicate that the consensus sequence shown in Fig. 5 reflects a mixture of many versions that

differ from one another at a few positions (Rosenberg *et al.*, 1978; Graf *et al.*, 1979; Graf, 1979; Thayer *et al.*, 1981). No cloned unit has a sequence identical to that of the consensus sequence. The same situation may exist in the α-satellites of other species since "ladder" patterns are always observed upon exhaustive type "A" digestion (Rubin *et al.*, 1980a; Wu and Manuelidis, 1980; Donehower *et al.*, 1980; Musich *et al.*, 1980; Maio *et al.*, 1981a). At least some variants of the AGM sequence are within satellite domains. For example, as many as 10% of the monomer units contain an *Eco*RI cleavage site between residues 31 and 32 [inspection of the consensus sequence (Fig. 5) shows that a single base pair change at residue 31 (T→ G) in the consensus sequence yields an *Eco*RI site]. Digestion with *Eco*RI yields a typical ladder pattern indicating a clustering of units with *Eco*RI sites (Fittler, 1977). Two long α-satellite segments containing frequent *Eco*RI cleavage sites were recently isolated by molecular cloning (McCutchan *et al.*, 1982). Furthermore, essentially all the α-satellite in a single AGM chromosome isolated within a mouse–monkey somatic cell hybrid contains cleavage sites for *Eco*RI; a large percentage of this α-satellite also has cleavage sites for *Hae*III (T.N.H. Lee and M. F. Singer, unpublished) although this site is present in less than 3% of total α-satellite (Fittler, 1977; Rosenberg *et al.*, 1978; Graf *et al.*, 1979; Thayer *et al.*, 1981). Heretical size α-satellite fragments are generated in small amounts by *Hin*dIII digestion (Fittler, 1977; McCutchan *et al.*, 1982). Some tandem arrays of AGM α-satellite are interrupted (McCutchan *et al.*, 1982) by dispersed repeated sequences such as *Alu*-SINE or *Kpn*-LINE (see below) family members (Grimaldi *et al.*, 1981; Grimaldi and Singer, 1982). The frequency at which heretical repeat units and interruptions occur is not known.

2. *Classical Primate Satellites*

By classical, I refer to fractions separable by density gradient centrifugation, cryptic or not. In the African green monkey the overwhelming mass of the major classical satellite is α-satellite; three or four additional classical satellites have been noted but not characterized (Kurnit and Maio, 1974; Fittler, 1977).

The classical human satellites present a complex picture compounded by the fact that similarly named preparations are not necessarily identical from laboratory to laboratory (Macaya *et al.*, 1977; Manuelidis, 1978; Miklos and John, 1979; Mitchell *et al.*, 1979). Macaya and co-authors (1977) and Miklos and John (1979) valiantly summarized the confusions. Recently some clarity has begun to emerge, but evaluation of the literature remains difficult. In particular, data on chromosomal location obtained by *in situ* hybridization with different satellites are problematic because of the cross-hybridization of the fractions (see below) and the use of uncloned probes; it will not be summarized here. The story of the complex interrelations among the classical calf satellites (see above) perhaps hints at what we may, in the long run, expect to learn. Table II summarizes one

TABLE II
CLASSICAL HUMAN SATELLITES[a]

	ρ (gm/cm^3)	Estimated percentage of genome
I	1.687	0.2–0.5
II	1.693	1–2
III	1.697	1–3
IV	1.700	0.5–2

[a] Data summarized from Macaya *et al.* (1977) and Mitchell *et al.* (1979).

overview of the properties of the major classical human satellites I, II, III, and IV (Corneo *et al.*, 1968, 1970, 1971; Macaya *et al.*, 1977; Mitchell *et al.*, 1979). Together they represent only 2–5% of the genome although fractionation of human DNA on the basis of renaturation kinetics suggests that this may underestimate the total (Marx *et al.*, 1976; Schmid and Deininger, 1975). The percentages of G·C base pairs estimated from the densities and melting temperatures are at variance with one another (Macaya *et al.*, 1977; Mitchell *et al.*, 1979). Furthermore, Mitchell *et al.* (1979) have presented data indicating that the density fractions contain related sequences. Trace amounts of ^{32}P-labeled III or IV reassociate with identical kinetics in the presence of an excess of unlabeled III and the melting curves of the resulting labeled duplexes are indistinguishable. Satellites I and II each share common sequences with III (and thus presumably IV); ^{32}P-labeled II hybridizes with III, as does ^{32}P-labeled I. The III-like sequences in I and II are probably not the same. It is not surprising then that all the satellites hybridized *in situ* to the same regions of the same human chromosomes. The classical satellites do not tell the whole story either; substantial satellite sequence may be buried in the main density band of DNA (Corneo *et al.*, 1980).

Of all the classical human satellites, III has been best studied (assuming that what is called III is the same in each instance). It contains at least four different types of sequence as defined by restriction endonuclease products and the ability to cross hybridize. The extent to which the different sequence types are covalently linked together is not known.

One portion of satellite III is degraded to a ladder of bands $n \times 170$ base pairs in length by restriction endonuclease *Hae*III (Manuelidis, 1976; Bostock *et al.*, 1978; Mitchell *et al.*, 1979, 1981). Similar bands are produced from satellite II (Mitchell *et al.*, 1979). The repeat length of these units is immediately reminiscent of α-satellite type sequences and they may indeed be related. The uncloned 340-bp α fragment set (see Fig. 5) isolated from total human DNA after cleavage with *Eco*RI hybridizes with *Hae*III ladders produced by digestion of total

genomic human, chimpanzee, gorilla, and simiang DNA (Manuelidis, 1978; Maio *et al.*, 1981a). *Eco*RI digestion of satellite III itself yields a 340-bp fragment (Mitchell *et al.*, 1979). Confirmation of the relation between the *Eco*RI dimer and sequences in the *Hae*III ladder by hybridization with cloned probes is required. Using uncloned total satellite III as a probe and rodent human somatic cell hybrids that contained a limited number of identified human chromosomes Beauchamp *et al.* (1979) found that certain size classes of *Hae*III generated fragments typical of total satellite are missing on specific chromosomes.

A second type of sequence included in satellite III is constructed of imperfect tandem repeats of the sequence 5′-TTCCA-3′ (see sequence 2 and 3, Fig. 6) (Deininger *et al.*, 1981); one of these is localized to the Y-chromosome (see below). A third type contains interrupted imperfect repeats of 5′-TTCCA-3′; at least two quite different versions of this type have been characterized within cloned segments (see sequences 1 and 4, Fig. 6) (Cooke and Hindley, 1979; Deininger *et al.*, 1981). One of them is localized mainly to chromosome 1 (Cooke and Hindley, 1979). Since all these different cloned segments were chosen essentially at random, a very large number of different versions of the 5′-TTCCA-3′ sequence may occur, each in long tandem arrays that constitute domains. Satellite II may contain related sequences (Manuelidis, 1978; Mitchell *et al.*, 1979). Inspection of the α-satellite sequences in Fig. 5 reveals a marked frequency of sequences like 5′-TTCC-3′ or 5′-GGAA-3′ (equivalent to 5′-TTCC-3′ on the other strand). This may bespeak a relation between the satellite III sequences and α-satellite (Deininger *et al.*, 1981).

A fourth class of sequences included in satellite III occurs only in male DNA and thus resides on the Y-chromosome. Upon digestion with *Hae*III, *Eco*RI, or *Eco*RII, male DNA yields a 3.4-kbp fragment set that is not found in female DNA (Cooke and McKay, 1978; Cooke and Hindley, 1979; Bostock *et al.*, 1978; Kunkel *et al.*, 1979). This fragment set may account for 40% of the entire Y-chromosome. Experiments with both cloned and uncloned 3.4-kbp fragments show that at least 50% of the sequences within the 3.4 kbp set hybridize with satellite III segments that are also abundant in female DNA. The remainder are male specific. Both types of sequences may be linked together in the members of the set (Kunkel *et al.*, 1979). This suggests an interspersion of Y-specific and Y-nonspecific sequence elements in a domain with a 3.4-kbp repeat length. The Y-specific domain is located on the long arm of the Y-chromosome (Kunkel *et al.*, 1977; Bostock *et al.*, 1978). Other domains that include the Y-nonspecific portion of the Y-domains but not within a 3.4-kbp repeat unit are represented in different amounts and different arrangements (domains) in various autosomes (Cooke and McKay, 1978). The properties of two cloned segments picked at random from a population of repeated sequences confirm the earlier analyses (Deininger *et al.*, 1981). One of the two hybridizes only to male DNA (sequence 2, Fig. 6) while the other hybridizes both to the male 3.4-kbp fragment set and to

1. Located on chromosome 1

5' AATTCATTTGAAGACAATTCCATTCAATACCAATTGATGATGGTTATTTTTGATTCCATTTGATGATGATTACATTCCAT

TTCATCATAATTCCATTCGATTCCACTCGAGATTCCATTCGATTCCATTCAA.....

.....CGAATGAATGAGTCCATCCATTTCAATTTCATGATAATTCCATTCGTTTCAATTCGATGGTGTTTCCATTC

GATT.....TTCATTCGATTCATTTGATGATGATTCATGCGCGATTCA

TTAGATGATGACCCCTTTCATTTCCATTCAATGGAGGATTCCATTCGGTTCCAT.....3'

2. Located on Y chromosome

5' TTAATTCCATTCCATTCCATTCCATTCCATTCCATTCCATTCCATTCCGTTCCATTCCATTCGTGTTGATTCCAT

TCCATTCCATTCCACTCCATTCCAATCCATTACATTCCACTCGGGTTGATTCCATTCCTTTCCATTCCAATCCATTCCAT

TCCATTCC 3'

3. Located on Y chromosome and elsewhere

5' CATTCAAGAAAGTTCCATTCCAGTCCATTCCCTTCGATTCCATTCCATTCGATTCCATTCTACTCGATTCCAATCTTGTC

CATTCCGTTGCATTCCATTCTATTCCATTCCATTGCATTGCATTCCATTCCATTTGATTACATTCCATTATATTCCATTC

CATTCA 3'

4. Location unknown

5' TTATTCCATTAGATTCCATTCGATGATGATTCCATTCGATTCCATTTGATGATTGCATTCTATTTCATTTGATGATGATT

CCATTCGAGTCCACTCGATGATTCCATTCGAGTCCATTCATTGATTCCATCCGATTTCATTGGATGATGACTCCATTCGA

GTCCATTCGATGATTCCACTCGATTCCATTAGATGATTCCATTGGAGTCCATTTGATTGTTCCATTCGATTCCATTCGAT

TCCT 3'

FIG. 6. Related but variant domains in human satellite III. (1) From Cooke and Hindley (1979); (2, 3, and 4) from Deininger *et al.* (1981).

female DNA (sequence 3, Fig. 6). Both are constructed of regular repeats of 5′-TTCCA-3′ and variants thereof; the differing variations in the two cloned segments must explain the specificity of hybridization. Note that a total of four segments, all included in satellite III and all built from variations of 5′-TTCCA-3′, have been described (Cooke and Hindley, 1979; Deininger *et al.*, 1981); their

structures are summarized on Fig. 6. In the absence of primary sequence data, the different hybridization specificities of the fragments would have obscured their similarities.

Little is known about satellite structure in other primates except that α-type sequences are ubiquitous (see above) and that satellite sequences in anthropoid apes are similar to one another (Deininger and Schmid, 1976b; Marx *et al.*, 1979; Mitchell *et al.*, 1981).

E. Marsupials

Macropus rufogriseus, the red necked wallaby, contains a satellite that can be isolated on CsCl gradients containing actinomycin D. The bulk of the satellite is degraded by *Bam*HI to sets of fragments that are from 1 to 5 times 2.5 kilobase pairs in length (Dunsmuir, 1976; Dennis *et al.*, 1980) and the repetition frequency is 5×10^5 in the genome. The 2500-bp repeat may be internally repetitive with a periodicity of about 300 bp. Secondary digests with other enzymes of both total satellite DNA and monomer units allow a division into distinct variant subsets of the basic structure. The variant subsets are clustered in eight domains, A through H, the characteristic structures of which are shown on Fig. 7. Within each domain, occasional repeat units have altered sequences that result in ladders of fragments with different enzymes. A model for the evolution of the satellite (see arrows on Fig. 7) was constructed on the principle that a restriction site common to the units in several domains was present in a common progenitor.

F. Carnivora

Centromeric C-banding is rare or nonexistent in members of the order Carnivora and early attempts to demonstrate a density satellite in *Felis catus* (Pathak and Wurster-Hill, 1977) failed. Matthews *et al.* (1980) have now reported that a cryptic satellite is revealed by density gradient centrifugation in the presence of netropsin. Analysis by reassociation kinetics after purification of the satellite by repeated centrifugation indicated a $C_0t_{\frac{1}{2}}$ of 7×10^{-4} mole/seconds/liter.

G. Comparisons between Satellite Sequences

Brutlag (1980) has recorded some statistically significant similarities between short sequences in the complex satellites of various species including *Drosophila* and several mammals. Thus residues 31 through 53 in the AGM α-satellite consensus sequence (Fig. 5, #1) are homologous to a portion of the 1.706D bovine satellite and residues 73 through 98 of the human sequence (Fig. 5, #4a) are homologous to a portion of the rat consensus sequence (Fig. 2b). Mouse satellite can be added to the list. Residues between 7 and 17 of the AGM consensus

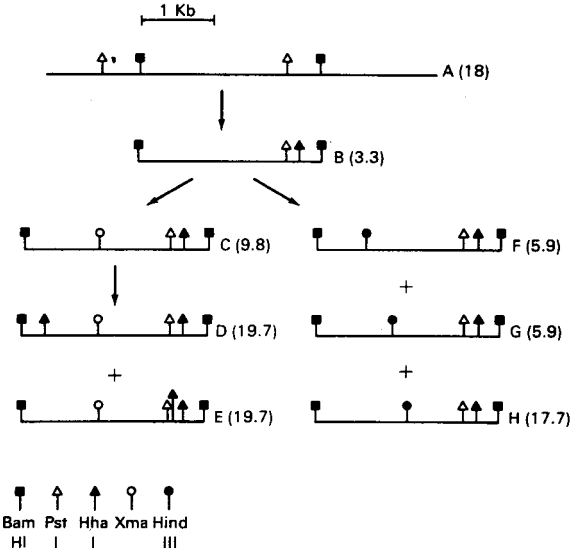

FIG. 7. Domains (A through H) within the red necked wallaby satellite (Dennis *et al.*, 1980). The numbers in parentheses represent the approximate percentage of total satellite in each domain.

(5′CTGAGAAACTG) are more than 87% homologous to multiple regions of the mouse consensus and are themselves repeated in part in 3 other positions of the AGM consensus. Further, the sequence is related to the progenitor nonanucleotides proposed for the mouse satellite (see above). Also, there is a significant similarity (93%) between residues 91 to 103 of the AGM sequence and 123 to 136 of the mouse. Unfortunately, statistical significance is neither necessary nor sufficient for biological importance. Questions about common and functionally significant satellite sequences in a range of species await new approaches.

H. Chromosomal Location

Are specific satellites or domains located on particular chromosomes? This important question remains largely unanswered except for the few instances already mentioned. Yet it is an important question for speculating about satellite evolution and is relevant to certain proposals about satellite function. Early experiments that used radioactive classical satellite preparations or RNA copies thereof as probes for *in situ* hybridization to chromosomes are now suspect not only because of possible contamination with other sequences but also because of the cross-hybridization of related but distinct satellites. In some instances, experiments with cloned and characterized probes should afford definitive data. But where different domains cross-hybridize (e.g., in the African green monkey and

the wallaby) chromosome specificity will have to be determined with isolated chromosomes. Somatic cell hybrids are a stable source of some individual chromosomes and this approach is being used (as reviewed above). Furthermore, direct methods for the isolation of specific chromosomes are being improved (Davies *et al.*, 1981).

I. The Question of Satellite Function

There is no experimental evidence regarding the function of mammalian satellite DNA. Speculation about function presently rests on the universal (but not necessarily exclusive) association of satellite DNA with heterochromatin (Brutlag, 1980; John and Miklos, 1979). However, heterochromatin function is itself not understood at the molecular level and most of the relevant information is limited to simpler eukaryotes that are amenable to either genetic or cytogenetic analysis or both. Reviews of earlier experiments and speculations are available (Bostock, 1980; Miklos and John, 1979; Hilliker *et al.*, 1980; Brutlag, 1980; John and Miklos, 1979; Orgel and Crick, 1980; Doolittle and Sapienza, 1980). The available data are most consistent with the idea that satellite functions in germ line processes (Bostock, 1980). The following two experimental systems may prove relevant to the question of satellite function.

1. *Transcription*

Transcription of satellite DNA occurs on specific loops in oocyte lampbrush chromosomes of newts (Varley *et al.*, 1980a,b; Diaz *et al.*, 1981; Gall *et al.*, 1981). Within these genomes, the satellite occurs both in centromeric regions and at the chromosomal regions corresponding to the loops. In *Notophthalmus viridescens,* hundreds of 9-kbp-long clusters of histone genes are each embedded within variable lengths of satellite sequence (estimated to be up to 100 kbp long) at the loops. Transcription may initiate at promotor sites preceding the histone genes and then proceed through a cluster continuing into the downstream satellite sequences (Varley *et al.*, 1980b; Diaz *et al.*, 1981). This is viewed as a failure of normal transcription termination but whether that failure is functionally significant remains to be learned.

There is one report that rat satellite sequences are transcribed in HTC rat tissue culture cells (Sealy *et al.*, 1981). This is of great interest, since earlier experiments with less sensitive techniques showed no transcription in a variety of species of organisms and cell lines.

2. *Centromeric Function*

The structure and function of yeast centromeres are now being investigated at the molecular level (Clarke and Carbon, 1980a,b; Hsiao and Carbon, 1981; Fitzgerald-Hayes *et al.*, 1982; Clarke *et al.*, 1981). The ability to couple precise

information on DNA structure with detailed genetic analysis makes this an especially powerful system. Using recombinant DNA techniques small circular DNA molecules have been constructed that contain selectable yeast genetic markers (e.g., genes required for synthesis of a particular amino acid) and a yeast DNA replicator sequence. Such plasmids replicate autonomously in yeast cells but are mitotically unstable and are rapidly lost from the population in the absence of dependency on the marker gene. The addition of a DNA segment spanning a centromeric region imparts both meiotic and mitotic stability to the plasmid. It then behaves like a true minichromosome, and becomes a stable component of the cells even in the absence of the selective pressure. During crosses involving two minichromosomes bearing identical centromeres but distinguishable marker genes, the two can apparently pair, moving to opposite poles in the first meiotic division (Clarke *et al.*, 1981). The centromeres of yeast chromosomes 3 (Clarke and Carbon, 1980a) and 11 (Fitzgerald-Hayes *et al.*, 1982) have been characterized. Yeast does not contain satellite sequences and there are no markedly repetitious segments in the defined centromeric regions. The essential regions of centromeres 3 and 11 are not more than 627 and 900 base pairs in length, respectively, and do not cross-hybridize with one another. The centromere 3 sequence is not repeated elsewhere in the genome. Yet *both* contain long A·T-rich regions flanked by short (14 bp or less) regions of homology. One of the latter is reported to display limited sequence homology with the 1.688 gm/cm^3 satellite DNA of *Drosophila melanogaster* (Clarke *et al.*, 1981; Hsieh and Brutlag, 1979a).

J. Amplification of Satellites

Most satellites are restricted to one or a closely related group of species and thus many appear to be of relatively recent origin. This in turn suggests that satellites change in evolution more than many other genomic regions. Compare, for example, the very close similarities between the β-globin regions (Barrie *et al.*, 1981) and the *Alu* families (see below) of the primate genomes with their strikingly different though related satellite domains. Large scale deletions and amplifications were required to arrive at present day satellites. The similarities between satellites of related species led to the "library" hypothesis regarding satellite amplification and deletion (Salser *et al.*, 1976; Fry and Salser, 1977); a common library (or set) of related sequences is available within related species and different members of the library are amplified to differing extents in the several organisms. At present there is no reason to assume a preexisting or common library. It is equally probable that newly divergent sequences were and are candidates for amplification (Gillespie *et al.*, 1980).

Mammalian genomic sequences can be amplified. The organization of the multiple globin genes (Fritsch *et al.*, 1980, 1981) demonstrates this phenomenon

in evolutionary time. The amplification of certain eukaryotic genes also occurs in experimental time. For example, clones of murine tissue culture cells containing many copies of the gene for dihydrofolic acid reductase (Alt *et al.*, 1978; Dolnick *et al.*, 1979) or of the complex of genes responsible for synthesis of aspartyl transcarbamylase (Wahl *et al.*, 1979) are readily obtained under appropriate selective pressure. In at least one instance, amplification of the mouse dihydrofolate reductase gene was accompanied by amplification of mouse satellite (Bostock and Clark, 1980). Once amplified, extra copies of these genes are readily, but not necessarily, lost when selective pressure is removed. Several mechanisms for amplification of tandem repeats have been discussed (Tartof, 1975; Botchan *et al.*, 1978; Schimke *et al.*, 1980; Smith, 1976; Kurnit, 1979; Baltimore, 1981). Among these, unequal crossing-over (Smith, 1976) is a relatively simple mechanism that is consistent with available data and has been demonstrated experimentally with the tandemly repeated ribosomal RNA genes of yeast (Petes, 1980; Szostak and Wu, 1980). In brief, homologous crossing-over between nonallelic repeat units in a tandem array (on sister chromatids or homologous chromosomes or homologous regions on nonhomologous chromosomes) yields an unequal number of repeat units in the products of recombination. Besides resulting in reciprocal amplification and deletion, unequal crossing-over provides for the maintenance of sequence homogeneity in tandem arrays (Smith, 1976). It also predicts that repeat units near the ends of tandem arrays will be more divergent than those near the center (Smith, 1976; Brutlag, 1980), as found with the AGM satellite (McCutchan *et al.*, 1982). Another mechanism for assuring homogeneity, called gene conversion, is described in Section V,D. This mechanism may be relevant to the fact that satellites on different chromosomes are characteristic of the species, even though they may be distinct domains (compare for example the a-a-a-a type structure of African green monkeys with the a-b-a-b-a-b of baboons).

Amplified genes, besides being subject to deletion, can follow at least three additional courses. First, they may be used essentially as such, as are the two adult α-globin genes. Second, they may evolve independently into either related functional genes, as for example, the embryonic and fetal globin genes, or into two quite different genes. Finally, a duplicated gene may be nonfunctional, as for example, the pseudogenes in the globin gene clusters. Presumably, different selective conditions favor one or the other course.

In this context, the situation with satellites is very striking. Is there in the sequences an inherent tendency toward amplification greater than that of, say, the dihydrofolic acid reductase gene or is there positive or negative selective pressure that operates to maintain multiple copies? Is there a selective force that dictates a minimal number of copies? Is there a copy number at which additional amplification is rejected? Is the location at centromeric and telomeric positions influential in determining copy number? It is not sufficient to argue that these

DNA segments exist only for their own replication (Orgel and Crick, 1980; Doolittle and Sapienza, 1980), without asking how they get away with this profligate reproduction while other sequences do not. What distinguishes a sequence that is tolerable in millions of copies from all the other genomic segments that are not? Is it the inability to be transcribed? Do the satellites provide a particular environment for genes or special sequences buried within them? The minimum requirements for centromere function in mitosis and meiosis can be supplied without millions of tandem repeats, as is evident from the recent work on yeast centromeres described above. It is conceivable that only one or a few copies of the satellite sequences are essential to cell function and that, unlike many other genomic segments, extra copies are tolerable safeguards and collect by virtue of amplification and the lack of negative selective pressure. This hypothesis provides a middle ground between the bold suggestion that satellite has no function at all (Orgel and Crick, 1980; Doolittle and Sapienza, 1980) and the well grounded assumption that biological systems are efficient.

Another set of questions arises in relation to the heterochromatic location of satellite. Is late replication in the cell cycle a consequence of heterochromatin structure or is it dictated by the sequences themselves? Does heterochromatin structure supply a useful positive function or does it represent a defensive mechanism whose chief advantage is to protect the genome from the effects of satellite sequences while still permitting functioning of those sequences at appropriate times or places? The latter idea is consistent with other examples of selective heterochromatin formation, as in X-chromosome inactivation.

IV. Interspersed Repeated Sequences

A. Emerging Concepts

Interspersion analysis—the measurement of the percentage of denatured DNA that registers as double-stranded on hydroxyapatite at a fixed low C_0t value as a function of chain length (Davidson *et al.*, 1973)—demonstrated that some highly repeated sequence families are dispersed among single-copy sequences in eukaryotic genomes. Two types of interspersion patterns were discerned. The *Drosophila* pattern (Manning *et al.*, 1975) is characterized by families of repeat units several kilobase pairs in length separated by tens of kilobase pairs of single-copy sequence. This arrangement has been confirmed by molecular analysis and several repeated elements in *Drosophila* are known to be mobile units (for review see Calos and Miller, 1980; Spradling and Rubin, 1981). The *Xenopus* pattern (Davidson *et al.*, 1973) is characterized by families of repeat units a few hundred base pairs in length separated by up to a few thousand base pairs of single-copy sequence. Very few organisms and not even all insects have

the *Drosophila* type pattern but many, including most mammals, are of the *Xenopus* type. Several organisms including chickens (Musti *et al.*, 1981) show interspersion patterns between the two extremes. And within the *Drosophila* genome another kind of arrangement, scrambled clusters of different short, moderately repeated sequence elements, occurs (Wensink *et al.*, 1979).

Substantial analysis at the molecular level has now revealed additional complexity. Two different classes of mammalian interspersed and highly repeated sequences have thus far been distinguished on the basis of size and relative abundance. Dispersed families with unit lengths under 500 base pairs are found in as many as hundreds of thousands of copies; they are here termed SINES, for short interspersed repeated sequences. In some organisms (e.g., humans) the copy number of the major SINE family is higher than the copy number of some satellite sequences, thus emphasizing the fact that satellite is not always the most rapidly reannealing portion of a genome. Other dispersed mammalian families are several kilobase pairs in length and occur on the order of 10^4 times (or fewer); they are here termed LINES, for long interspersed repeated sequences.

B. Short Interspersed Repeated Sequences (SINES)

SINE families often contain more than 10^4 member sequences although smaller families also exist. The many members in a family are similar enough to hybridize to one another under stringent conditions, but are not identical (Rinehart *et al.*, 1981). It is quite possible that different families in a particular species are related to one another. Although a variety of functions have been suggested (Britten and Davidson, 1969; Davidson and Britten, 1979; Jelinek *et al.*, 1980) the physiological significance of these sequences remains unknown. The strongest available arguments for assuming a functional importance for at least some members of these families are their marked conservation and the fact that they are transcribed. The characterized families of rodent and primate SINES share common features but are, at the same time, quite different.

The repeat units occur in both possible directions within genomic segments; the significance of the direction, if any, is unclear. This arrangement accounts for many if not most of the abundant short inverted repeats previously described in mammalian genomes and is reflected in nuclear RNA molecules that also contain inverted copies of family members. Transcripts of family members are most abundant in heterogeneous nuclear RNA but also occur in cytoplasmic RNA.

1. *Human Alu Family*

The best characterized mammalian SINE family is the *Alu* family of the human genome. The presence of a large set of cross-hybridizing short DNA sequences dispersed throughout most of the human genome and occurring frequently in both possible orientations was demonstrated initially by renaturation kinetics, hyd-

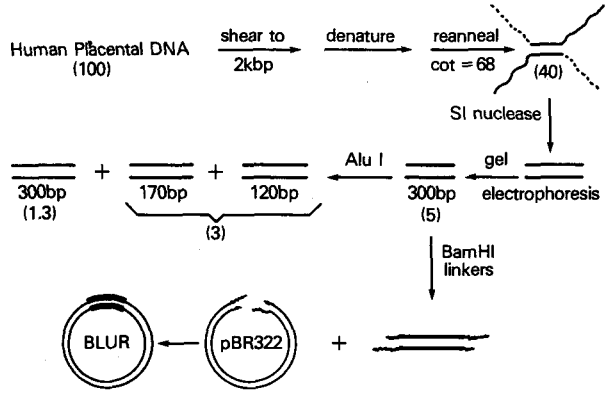

FIG. 8. The isolation and cloning of the human *Alu* family (Houck *et al.*, 1979). The cloning vector was the *E. coli* plasmid pBR322. The numbers in parentheses indicate the approximate percentage of total DNA in each fraction.

roxyapatite separation, interspersion analysis, and electron microscopy (Schmid and Deininger, 1975; Deininger and Schmid, 1976a). *Alu* family sequences were first isolated (Fig. 8) (Houck *et al.*, 1979) by renaturing sheared (2 kbp) denatured human placental DNA to a C_0t of 68 and removing single-stranded tails with S1 nuclease. Analysis of the size of the double-stranded products by gel electrophoresis revealed a broad range of sizes but about 5% of the products were 300 bp in length. More than 60% of the 300-bp-long set of duplexes was cleaved by *Alu*I into two fragments of about 170 and 120 bp in length, respectively (Houck *et al.*, 1979): thus the name, *Alu* family. More precise measurements now put the reiteration frequency of the family close to 3×10^5 copies or about 3% of the genome (Rinehart *et al.*, 1981). Other families of SINES may also be interspersed in the human genome (Fritsch *et al.*, 1981; Deininger *et al.*, 1981), but it seems unlikely that any approach the copy number of the *Alu* family (Rinehart *et al.*, 1981).

Direct evidence for the distribution of *Alu* sequences comes from studies with long cloned segments of the human genome. Recombinant libraries that comprise the bulk of the human genome divided into segments 15–20 kbp long, each inserted into a λ-bacteriophage vector, are available (Maniatis *et al.*, 1978; Lawn *et al.*, 1978). Screening of individual phage in a library by hybridization with a cloned *Alu* segment indicates that more than 90% of the human segments hybridize with *Alu* (Tashima *et al.*, 1981). Many phage that were selected from a library for other reasons also contain *Alu* sequences. There is, for example, an *Alu* sequence some 6 kbp downstream from the 3′ end of the insulin gene (Bell *et al.*, 1980); it is the only *Alu* sequence within about 19 kbp. There are 8 *Alu* sequences within the 65 kbp region that contains the entire cluster of β-globin

genes (Duncan *et al.*, 1979; Coggins *et al.*, 1980; Baralle *et al.*, 1980; Fritsch *et al.*, 1980). Eight cloned segments randomly selected from a human library have a minimum of 9 *Alus* in a total of 88 kbp (Pan *et al.*, 1981). And 3 *Alu* segments are within a 15 kbp segment that contains DNA homologous to the transforming gene of simian sarcoma virus (Dalla-Favera *et al.*, 1981); all three are within putative intervening sequences (introns). Altogether this represents 21 *Alu* sequences in a total of 187 kbp of DNA or one *Alu* sequence every 9 kbp. If a similar distribution occurs in all the library segments then there are an average of 1–2 *Alu* family members in each of the 90% that hybridized, approximating the 3×10^5 copies estimated by other means.

The primary nucleotide sequence of 11 randomly selected and cloned members of the *Alu* family is known (Rubin *et al.*, 1980b; Deininger *et al.*, 1981; Pan *et al.*, 1981). Ten of the 11 are part of a group referred to as BLUR (Rubin *et al.*, 1980b; Deininger *et al.*, 1981) and were selected from the 300-bp-long fragments obtained from the total genome (Fig. 8). These 10 sequences were used to construct a consensus sequence (see Fig. 9,1). Although the individual BLUR sequences differ from each other (and from the consensus) both by insertions, deletions, transitions, and transversions, they are, on the average, 88% alike. The alterations are spread around the sequence thus accounting for the strong cross-hybridizations. In addition, the sequence of four *Alu* family members from known positions in the human genome have been determined: one near the insulin gene (Bell *et al.*, 1980) and one each preceding the ϵ-globin gene (Baralle *et al.*, 1980), the Gγ-globin, and the δ-globin genes (Duncan *et al.*, 1981). All are approximately 300 bp long and are about 80% homologous to the consensus.

Certain very striking features of *Alu* family members have emerged from the sequence analysis. First, the 300 bp unit is essentially a head-to-tail dimer of a sequence about 135 bp long. One-half of the internally repeated sequence is interrupted by about 30 extra bp (approximately residues 218 to 250, Fig. 9,1). Each half is terminated by a stretch of A·T base pairs though this is generally most striking toward the right end (as written, Fig. 9,1). Within each half is a 14-bp segment (residues 41 to 53 and 176 to 188, in Fig. 9,1) which is similar to a sequence around the origin of replication of papovaviruses (Jelinek *et al.*, 1980).

Short direct repeats flank the *Alu* sequence in each instance studied (Fig. 10); the BLUR clones do not include flanking sequences because of the way in which they were isolated. These flanking repeats are not part of *Alu* itself and are different in each case examined both in length and structure (Fig. 10). A similar direct repeat has been found in a sequenced *Alu* segment from the African green monkey genome (see below and Fig. 10).

2. *Transcription of Human Alu Sequences*

The percentage of cytoplasmic RNA that hybridizes to *Alu* is at least 10-fold less than that found in nuclear RNA. Heterogeneous nuclear RNA (hnRNA)

```
1  GCT GGGCGTG  G     TGGCTCACA  CCTGTAATCC  CAGCACTTTG  GGAGGCCGAG  GTGGGTGGAT
2  GCC GGGCGGG  A     TGGCTCATG  CCTGTAATCC  CAGCACTTTG  GGAGTTCGAG  GCGGCAGGAT
5    CGGGGC TG  GAGAGATGGCTCAGC  GGTTAAGAGC    GCCC       GACTGCCTT
6     GAGGC TG  GAGAGATGGCTC GA  GGTTAAGAGC     ACCA       ACTGCTGT

1  CACCTGAGGT  CAGGAGTTCA       AGACC  AGCCT  GGCCAACATG  GTGAAACCCC  GTCTCTACTA
2  CACCTGAGGT  CGGGAGTTCG       AAACC  TGC T  GGCCAACATG  GCGAAACCCC  GTCTCTACTA
5    CCAGAGGT  CATGAGTTCA       ATTCCCAGC     AACCA CATG  GTGGCTCACA  ACCATCTGTA
6   TCCAGAGGT  CCTGAGTTCA       ATTCCCAGC     AACCA CATG  GTGGCTCATA  ACAATCTATA

5  AAGAGATCTG  ATGCCCTCTT       CTGGTGTATC    TGAAGACAGC  TACAGTGTAC  TTATATATAA
6  ATGAGATCTG  GTGCCCTCTT       CTGGTGTGCA    GATATATATG  GAAGCAGAAT  GTTGTATACAT

1  AAAATACAAA  AATTA
2  AAAATACAAA  AATTA
5  TAAATAAATA  AATCTTTAAA       AAAAACAAAA    CAAAAACAAA  AACAAAA
6    AATAAATA  AATAAAATCT       TAAAAAAA

1  GCCGG       GCGT  GGTGGC     GCGCGCCTGT    AATCCCAGCT  ACTCGGGAGG  CTGAGGCAGG
2  GCCG        GTGT  GGTGGC     GCATGCCTGT    AGTCCCGGCT  ACTCGGGAG   CTGAGGCAGG
3   CCGG       GCAT  GGTGGT     GCATGCCTTT    AATCCCAGC   ACTCGGGAGG  CAGAGGCAGG
4   CCAG       GCATT GGTGGT     ACACACCTTT    AGTCCCAGC   ACTCAGGAGG  CAGAGGCAGG
                        C

1  AGAATCGCTT  GAACCCAGGA       GGTGGAGGTT    GCAGTGAGCC  GAGATCGCGC  CACTGCACTC
2  AGAGTTGCTT  GAACCT GGA       GGTGGAGGTT    TCAGTGAGAC  AAGATCACAT  CACTGCAC
3  CGGATTTCT   GAGTTCGAGG                                                    C
4  AGGATCACTT  GAGTTCAAGA       G                                           C
                     C
                     T

1  CAGCCTGGGC                                                      AACAGAGCGA
2  CAGCCTGGGC                                                      ACAGAGCAA
3  CAGCCTGGTC  TTC AGAGT          GAGTTCC     AGGACACCAG  GGCTA     CAGAGAAA
4  CAGCCTGGTC  TACCAGAGTT       CCTGAGTT C    AAGCCA      GGCTAT    ACAGAGAAA
                     T
                     A

1  GACTCCATCT  C                AAAAAAAAA     AAAAAAAAAA  AAA
2  CACTCTA     AGAGAGAG         AAAAAAAAA     ACCCACAAAA  AAA
3  CCCTGTCT  - A rich
4  CCCTGTCT    [(A)nN]1
```

FIG. 9. Comparison of SINES. (1) is the human *Alu* consensus sequence (Deininger *et al.*, 1981); (2) is a cloned monkey *Alu* sequence (Grimaldi *et al.*, 1981); (3) is the mouse B1 consensus sequence (Krayev *et al.*, 1980); (4) is the Chinese hamster consensus sequence (Haynes *et al.*, 1981); (5) is the second half of the dimeric SINE found in the intervening sequence of the rat growth hormone gene (the sequence is reported in Page *et al.*, 1981 and the homologies were found using the computer program of Queen and Korn, 1980); (6) is a Chinese hamster sequence (Haynes *et al.*, 1981) which resembles the rat sequence shown in line 5. The marks above the human *Alu* sequence (1) indicate every tenth base pair in that sequence. Only one DNA strand is shown. Each base in each sequence is shown, in order, in the 5′ (upper left) to 3′ (lower right) direction. In sequence (4) frequent variations in the consensus sequence are noted under the line.

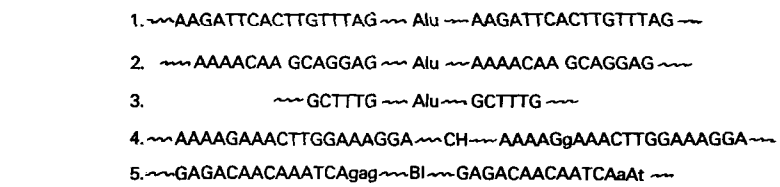

FIG. 10. Direct repeated sequences surrounding members of the primate and rodent SINE families. Selected examples are given. (1) *Alu* 5′ to human Gγ-globin gene (Duncan *et al.*, 1981); (2) *Alu* 3′ to human insulin gene (Bell *et al.*, 1980); (3) a monkey *Alu* (Grimaldi *et al.*, 1981); (4) a Chinese hamster SINE (Haynes *et al.*, 1981); (5) a mouse B1 sequence (Krayev *et al.*, 1980). Imperfections in the repeats are indicated by lower case letters. Other examples may be found in the following references: Elder *et al.* (1981), Baralle *et al.* (1980), Haynes *et al.* (1981), Page *et al.* (1981), and Grimaldi and Singer (1982).

contains high levels of *Alu* transcripts in molecules ranging from 100 to more than 5000 nucleotides in length (Federoff *et al.*, 1977; Jelinek *et al.*, 1978, 1980; Elder *et al.*, 1981). At least some of the transcribed *Alu* sequences are found in intra- and intermolecular RNA·RNA duplexes, presumably because of inverted orientations and the transcription of both strands of *Alu* family members. *Alu* transcripts are found in both polyadenylated and nonpolyadenylated cytoplasmic RNA (Pan *et al.*, 1981; Elder *et al.*, 1981; Jelinek *et al.*, 1978; Tashima *et al.*, 1981). The presence of *Alu* transcripts in polyadenylated RNA cannot be taken as evidence that the RNAs are polyadenylated by posttranscriptional modification, as are most eukaryote messenger RNAs, because the long terminal A stretches in the *Alu* sequence may account for the binding of these molecules to the oligo(dT)-cellulose used in standard separations (Elder *et al.*, 1981). A discrete 7 S cytoplasmic RNA from HeLa cells (uninfected) hybridizes to *Alu* DNA and this is the only abundant HeLa cytoplasmic RNA that hybridizes with *Alu* (Weiner, 1980). Analysis of hybrids formed between 7 S RNA and cloned *Alu* DNA indicates imperfect duplexes suggesting (Weiner, 1980) that 7 S RNA may be encoded by an unusual subset of the *Alu* family.

Many but not all *Alu* sequences in cloned human DNA fragments are transcribed *in vitro* by RNA polymerase III (Duncan *et al.*, 1979; Elder *et al.*, 1981). The structure of the transcripts is consistent with initiation at the start of the *Alu* sequence and termination beyond the end of *Alu* and the RNA has the sequence of the DNA strand in Fig. 9,1. Transcriptional control regions are generally present internal to the coding sequences of genes transcribed by RNA polymerase III (reviewed by Duncan *et al.*, 1981). The *Alu* consensus sequence includes a segment 5′-GAGTTCPuAGACC-3′ (at about position 75) that is similar to a sequence noted by Fowlkes and Shenk (1980) (5′-GGGTTCGANNCC-3′) as important in control of transcription initiation by RNA polymerase III. This sequence does not appear clearly in the second half of the *Alu* consensus; the

analogous position is interrupted by the extra base pairs. Similarly, termination of RNA polymerase III transcription of *Alu* occurs at a run of 4 or more T residues, as with other polymerase III templates (Duncan *et al.*, 1981). There is evidence suggesting that 7 S RNA is synthesized *in vivo* by RNA polymerase III (reviewed by Weiner, 1980; Zieve, 1981).

3. *Alu-like Sequences in Other Primates*

The bonnet monkey (Houck and Schmid, 1981) and African green monkey (Dhruva *et al.*, 1980; Grimaldi *et al.*, 1981) each has an abundant family of SINES closely similar to human *Alu* in copy number and sequence. The galago (*Galago crassicaudatus*) has a distantly related family of unknown copy number (Houck and Schmid, 1981). About 75% of recombinant phage representing a library of the African green monkey genome hybridize with a cloned human *Alu* sequence, corresponding to a minimum of 1.8×10^5 copies per haploid genome. Further, a group of 4 phage, chosen from the library essentially at random with reference to *Alu*, contain together at least 8 copies of *Alu* or an average of one every 8 kbp of DNA. The sequence of one of the monkey *Alu* segments was determined (Fig. 9,2); it is 84% homologous to the human *Alu* consensus sequence and is flanked by short direct repeated segments (Grimaldi *et al.*, 1981). Among the cloned monkey DNA segments were some in which the *Alu* sequence abutted or interrupted African green monkey α-satellite sequences (see Sections III,D and IV,B,6).

4. *SINES in Rodents*

One predominant family of SINES—the *Alu* family—have been described in primates while rodent genomes are known to contain several distinct SINE families of relative similar abundance. This is not to say that the *Alu* sequences are the only SINES in primate genomes; others that may be present on the order of 10^4 times have already been detected (Deininger *et al.*, 1981; Fritsch *et al.*, 1981).

Mouse heterogeneous nuclear RNA forms both inter- and intramolecular double-stranded regions (reviewed by Georgiev *et al.*, 1981). The double-stranded regions were isolated from intramolecular duplexes by digesting away single strands with ribonucleases T1 and A and the resulting RNA fragments were used to detect molecular clones of genomic DNA that hybridize with the RNA probes. Two distinct families of SINES termed B1 and B2 were isolated. Members of the two families each occur about 10^5 times scattered throughout the genome (Kramerov *et al.*, 1979; Georgiev *et al.*, 1981). These sequences, like *Alu*, occur in both orientations along the genome and the inverted configuration presumably accounts for much of the intramolecular double-stranded heterogeneous nuclear RNA. The sequences of 3 randomly chosen cloned members of the B1 family are known and were used to construct a consensus sequence

(Fig. 9,3) (Krayev *et al.*, 1980); the mismatch between any pair of the three is less than 8%. Two of the sequenced B1s represent separate units within a single 2.3-kbp-long cloned fragment of mouse genomic DNA.

The B1 consensus sequence is about 130 bp long followed by an A·T-rich region (A-rich at the 3′-end of the strand shown in Fig. 9,3). Beside the A-rich region there are other striking similarities between the mouse B1 sequence and the primate *Alu* sequences (Pan *et al.*, 1981). A 40-bp sequence that is repeated in each arm of *Alu* (residues 21 to 60 and 156 to 195, Fig. 9,1) is highly conserved in the B1 sequence (Fig. 9,3); this includes the 14-bp segment homologous to the origin of replication of papovaviruses. B1 sequences are also flanked by short direct repeats of DNA (see Fig. 10). Overall B1 can be seen as analogous to one half of the *Alu* dimer. However a human *Alu* sequence (clone BLUR 8) does not hybridize to mouse DNA under stringent conditions (Shih *et al.*, 1981).

The B2 family does not cross-hybridize with B1 (Kramerov *et al.*, 1979). Other evidence indicating multiple families of abundant mouse SINES is seen in the fact that three separate, noncross-hybridizing SINES were discovered interspersed among the 65-kbp-long mouse β-globin gene cluster; each of the three occurs twice or more within the segment (Haigwood *et al.*, 1981). The distribution of SINES in the β-globin gene cluster of two different *Mus musculus* strains (BALB/c and C57BL/10) appears to be essentially the same (Haigwood *et al.*, 1981). The three SINE families were labeled a, b, and c (Haigwood *et al.*, 1981) and it is not known if any of them are related to the previously described B1 and B2. Cloned members of the a and b families hybridized with essentially all the phage in a mouse genomic DNA library. Thus, there are at least 2×10^5 separate members in each of these families assuming that the inability to cross-hybridize extends to all members of each family. It is possible that mouse SINE family members are interspersed in mouse satellite (Haigwood *et al.*, 1981).

Five different SINE families occur at 20 positions within the 44-kbp segment containing the genes for the β-globins of rabbits (Shen and Maniatis, 1980; Fritsch *et al.*, 1981). Only one family, called C, hybridizes (weakly) to a human *Alu* sequence under nonstringent conditions. C sequences, like *Alu*, are transcribed by RNA polymerase III *in vitro*. Members of SINE families were found buried in long cloned segments of Chinese hamster DNA. The clones were selected because they hybridized with double-stranded heterogeneous nuclear RNA (Jelinek, 1978; Haynes *et al.*, 1981). Thirty-seven percent of the phage in a Chinese hamster genomic DNA library hybridized. A large proportion of the cloned segments had more than one hybridizing region and many of these occurred in opposite orientations. The sequences of 6 such regions were determined and a consensus derived from 5 of them (Haynes *et al.*, 1981) (Fig. 9, 4). The similarity with the mouse B1 consensus sequence is evident. One of the 6 (Fig. 9, 6) differs markedly from the others after the first 60 base pairs, but is, there-

fore, homologous to a rat sequence (see below). Like *Alu* and B1, the Chinese hamster SINES' are surrounded by short direct repeats of DNA (Fig. 10).

Much less is known about the situation in rats. Houck and Schmid (1981) reported the presence of abundant 300-bp-long segments in reannealed rat DNA that was treated with S1 as outlined in Fig. 8 for the original isolation of human *Alu* sequences. One repeated DNA segment occurs within an intron of the rat growth hormone gene and again after the end of the last exon in the gene (Page *et al.*, 1981). The sequence of the segment in the intron (Fig. 9, 5) includes two direct repeats of a segment about 200 base pairs long. This unit may represent a rat SINE family. It is, as noted above, similar to one Chinese hamster sequence (see Fig. 9, 5 and 6). A repetitive sequence has also been identified downstream from the rat insulin I gene (Bell *et al.*, 1980); it does not hybridize with a human *Alu* probe.

5. *Transcription of Rodent SINES*

As already described, the mouse B1 and B2 sequences are abundantly transcribed; copies of both strands of the sequence occur in heterogeneous nuclear RNA (Kramerov *et al.*, 1979). A unique version occurs in mouse cytoplasm as a discrete 4.5 S RNA (Harada and Kato, 1980); this sequence differs from the consensus sequence derived from three cloned members of the B1 family (Krayev *et al.*, 1980). The 4.5 S RNA hybridizes to both cytoplasmic and nuclear polyadenylated RNA (Jelinek and Leinwand, 1978). It is striking that the 4.5 S mouse sequence is more similar (3 differences out of about 90 nucleotides) to a 4.5 S cytoplasmic RNA of the Chinese hamster (see below) than are the mouse and Chinese hamster consensus DNA sequences to one another (see Fig. 9). One, or a few members of these families may be the functional genes for 4.5 S RNA in each of the species, and if so, that gene has been markedly conserved (Haynes *et al.*, 1981). Both the mouse and Chinese hamster 4.5 S RNAs are copies of the same DNA strand (sequence equivalent to the DNA strands in Fig. 9).

6. *Are SINES Functional?*

As a background, it is interesting to recall proposals suggesting that highly repeated dispersed sequences may be without function (Orgel and Crick, 1980; Doolittle and Sapienza, 1980) and also disagreement concerning those proposals (Cavalier-Smith, 1980; Dover, 1980; T. F. Smith, 1980; Orgel *et al.*, 1980; Dover and Doolittle, 1980). Specific functions that have been suggested include the control of gene expression (Britten and Davidson, 1969; Davidson and Britten, 1979; Jelinek *et al.*, 1980), perhaps by involvement of transcripts of SINES in the maturation of messenger RNA (Georgiev *et al.*, 1981; Jelinek *et al.*, 1980; Zieve, 1981; Lerner and Steitz, 1981), and service as origins of DNA replication (Jelinek *et al.*, 1980; Georgiev *et al.*, 1981). The cytoplasmic 4.5 S RNA of rodent cells and the 7 S RNA of human cells appear to represent abundant

transcripts of a particular member(s) of the respective SINE families. At least these SINE family members are likely to be functional genes. It is striking that the 4.5 S RNA of the hamster and mouse are identical in all but 3 of 90 nucleotides (Harada and Kato, 1980; Haynes *et al.*, 1981). They are both found in small nuclear ribonucleoprotein complexes, as well as cytoplasm and recent speculations regarding their function are reviewed by Zieve (1981) and Lerner and Steitz (1981). Although the 4.5 S RNAs have 5′-terminal triphosphates, consistent with *in vivo* transcription by RNA polymerase III, it is not clear to what extent RNA polymerase II transcripts that include SINES also contribute to homologous sequences in heterogeneous nuclear or cytoplasmic RNA.

Some clues regarding SINE function might be expected from a comparison of their distribution around corresponding genes in different organisms. Extensive data are available for the β-globin gene clusters of mouse (Haigwood *et al.*, 1981), rabbit (Shen and Maniatis, 1980; Hoeijmakers *et al.*, 1980; Fritsch *et al.*, 1981), and human (Fritsch *et al.*, 1981) and the α-globin cluster of humans (Fritsch *et al.*, 1981). Within each cluster are several coding regions, one or more pseudogenes, and various SINES. In the human β-cluster 5 genes are expressed in a timed sequence: ϵ in embryos, Aγ and Gγ during fetal development, and δ and β during adulthood. In the rabbit, two genes are expressed in embryos and one in adults, while in the mouse, two are expressed in adults and one in embryos (Jahn *et al.*, 1980). Both rat and human insulin genes (Bell *et al.*, 1980) have a SINE following the end of the last exon, although at different distances. There is no equivalent to the SINE within one intervening sequence of the rat growth hormone gene in a corresponding human gene (Page *et al.*, 1981). While some patterns can be discerned (e.g., in the rabbit the C-SINE family is clustered in the region of embryonic genes and in the human β-globin cluster no *Alu* occurs between simultaneously expressed genes) and may prove to be meaningful, no definitive statements can be made at this time.

Another interesting possibility is that some SINES are mobile or transposable sequence elements analogous to those described in yeast and *Drosophila* (reviewed by Calos and Miller, 1980; Spradling and Rubin, 1981). Mobile elements are flanked by short direct repeats of sequences that are not part of the element itself. In the case of characterized mobile elements the repeat lengths are characteristic of the element and constant, and the duplications reiterate the target site at which insertion occurred. Similarly, integrated DNA copies of RNA tumor virus genomes in birds and mammals are flanked by such short direct repeats (reviewed by Temin, 1980). These viral inserts as well as several well characterized mobile elements in bacteria, yeast, and *Drosophila* share other structural features. These include direct terminal repeats of several hundred base pairs at the two extremities of the movable element itself as well as a total length of several kilobase pairs. Also eukaryote mobile elements do not typically have the long terminal stretch of AT base pairs that is a common feature of mammalian SINES.

However, a recently discovered movable *Drosophila* element does have notable similarities to the mammalian SINES. This element, referred to as 101F (Dawid *et al.*, 1981), has no long terminal repeats, is flanked by short direct repeats of the target site, and contains a 3′-terminal poly(A) segment 18 nucleotides long. Unlike the mammalian SINES, 101F is about 4 kbp in length.

Several observations are consistent with the possibility that SINES are mobile in mammalian genomes, at least in evolutionary time. First, there is no counterpart in the human gene to the SINE within the intervening sequence of the rat growth hormone gene, although the genes themselves are likely to be homologous and the introns similarly placed. Thus, assuming a common ancestral sequence, the SINE was either lost during evolution of the human gene or acquired during evolution of the rat gene. Second, the very different distributions of SINES in the clusters of globin genes (comparing rodents with primates or the α- and β-clusters of humans) are consistent with mobility. Third, *Alu* sequences are found within African green monkey α-satellite segments (Grimaldi *et al.*, 1981). Because the *Alu* sequence is highly conserved and satellite was amplified after separation of the monkey and human evolutionary lines, the *Alu* sequence was probably acquired after amplification of the satellite. Finally, the direct repeats surrounding mammalian SINE segments (Fig. 10) are reminiscent of the target site duplications that flank mobile elements. Very recently the primary nucleotide sequence of an African green monkey α-satellite segment that is interrupted by an *Alu* family member was determined. A 13-bp-long repeat surrounds the *Alu* and this is a duplication of the α-satellite sequence at the point of interruption (Grimaldi and Singer, 1982).

Calos and Miller (1980) give an excellent summary of the possible significance of mobile elements in eukaryotes, emphasizing their potential for the generation of genetic diversity. The following additional point may be important, in view of the suggestions that highly repeated sequences have no function at all. A mobile element may generate diversity with a potential for selective advantage, but it can also generate disadvantage if it moves into an essential gene. Mutation by movable elements has been demonstrated in yeast (Roeder and Fink, 1980) and *Drosophila* (reviewed in Spradling and Rubin, 1981). The high frequency of mutation caused by the presence of large numbers of movable elements within a mammalian genome might have proven intolerable and been selected against, unless it was counterbalanced by some positive functional advantage.

Finally, the suggestion (Jelinek *et al.*, 1980; Georgiev *et al.*, 1981) that SINES may serve as origins for DNA replication should be considered. The basis for the suggestion is the presence in SINES of a short (14 bp) homology to a sequence associated with the origin of replication of murine and primate papovaviruses. Georgiev *et al.* (1981) describe some preliminary experiments that are consistent with this suggestion. However, in papovavirus genomes this region is part of a complex control region and may be involved in the control of

transcription as well as replication. Only additional experiments will resolve these questions.

C. Long Interspersed Repeated Sequences (LINES)

Dispersed repeated sequences several kilobase pairs in length (LINES) have been described in both rodents and primates. Several LINE families may eventually be found in each genome. As yet none is completely characterized.

1. *Primates*

Several groups of investigators appear to have discovered the same family of LINES, called here the *Kpn*-LINE family (Adams *et al.*, 1980; Kaufman *et al.*, 1980; Schmeckpeper *et al.*, 1981; Maio *et al.*, 1981b; Manuelidis, 1981b; Rogers, 1981; J. C. Rogers and C. Milliman, unpublished; G. Grimaldi and M. F. Singer, unpublished). A segment about 6.4 kbp long and estimated to be repeated 3 to 5 thousand times in the human genome was found 3 kbp downstream from the 3′ end of the human β-globin gene (Adams *et al.*, 1980; Kaufman *et al.*, 1980). This and several other somewhat divergent randomly selected members of the family were cloned from a human library. Portions of this LINE hybridize under stringent conditions with a cloned 2.8 kbp *Kpn*I fragment of African green monkey DNA; this fragment was found interrupting and abutting α-satellite and was independently identified as a portion of a monkey LINE found in thousands of copies in both human and monkey DNA (G. Grimaldi and M. F. Singer, unpublished). Other members of the monkey family diverge in such a manner that the sequences in the 2.8-kbp *Kpn*I fragment are divided into two *Kpn*I fragments, 1.2 and 1.5 kbp in length, respectively. The 2.8-kbp fragment hybridizes to both monkey and human genomic *Kpn*I fragments 1.2, 1.5 (1.6), 2.8, 3.4, and 4.6 kbp long; it also hybridizes to monkey and human *Hin*dIII segments 1.9 and 2.6 kbp long. One member of a set of abundant 1.9-kbp human *Hin*dIII fragments was cloned separately and shown to hybridize to *Kpn*I fragments of the same size (Rogers, 1980; J. C. Rogers and C. Milliman, unpublished). Independent work by Schmeckpeper *et al.* (1981) identified a LINE family that also hybridized with the same size classes of *Hin*dIII and *Kpn*I fragments from the human genome. The LINE family was located on both X and autosomal chromosomes. The *Kpn*-LINE family appears to be common among primates since digestion of total DNA from several species yields the distinctive set of *Kpn*I bands (Maio *et al.*, 1981b).

The dispersed organization of the *Kpn*I-LINE family is inferred from (1) its presence on human X and autosomal chromosomes (Schmeckpeper *et al.*, 1981; Manuelidis, 1981b), (2) its distribution in the phage of the monkey (G. Grimaldi and M. F. Singer, unpublished) and human (Adams *et al.*, 1980) libraries, and

(3) the absence of any "ladders" upon restriction endonuclease cleavage (all papers already mentioned).

Five percent of the bacteriophage comprising an African green monkey genomic library (McCutchan *et al.*, 1981) hybridized with the cloned 2.8 *Kpn*I fragment, consistent with about 12,000 dispersed copies of the sequence. Some estimates for the human genome (Maio *et al.*, 1981b; Rogers, 1981) are similar, although these numbers are markedly higher than the 3–5 thousand copies estimated by Adams *et al.* (1980). J. C. Rogers and C. Milliman (unpublished) have obtained data suggesting that the copy number of the *Kpn*-LINE family may vary from individual to individual.

It is already clear that there is extensive divergence among the members of the *Kpn*-LINE family although the various members cross-hybridize. Both the cloned human segments (Adams *et al.*, 1981) and cloned monkey segments (G. Grimaldi and M. F. Singer, unpublished) show restriction endonuclease site variability. Also, it is evident from digests of total genomic DNA hybridized to cloned probes that the abundance of different variants is very different in monkeys and in humans (G. Grimaldi and M. F. Singer, unpublished). Data suggesting that *Kpn*-LINE family members cross-hybridize with α-satellite (Maio *et al.*, 1981b) are puzzling since other investigations give no indication of such a relation (McCutchan *et al.*, 1981; G. Grimaldi and M. F. Singer, unpublished; Manuelidis, 1981b). It is probable that the uncloned α-satellite probes were contaminated with *Kpn*-LINES.

2. *Rodents*

Several groups of workers have obtained data on what appears to be a single abundant LINE family in mice, called here the *Eco*RI-LINE family. Between 1 and 3% of the *Mus musculus* genome is converted to a set of fragments 1.3 kbp in length upon digestion with *Eco*RI; this is equivalent to as many as 50,000 copies per haploid genome (Hörz *et al.*, 1974; Cheng and Schildkraut, 1980; Heller and Arnheim, 1980; Manuelidis, 1980; Brown and Dover, 1981; Meunier-Rotival *et al.*, 1981). The 1.3-kbp segment is frequently if not always contained within repeat units that are at least 5.6 kbp in length (Meunier-Rotival *et al.*, 1981; Brown and Dover, 1981). Although the sequence itself is conserved among the members of the family, subgroups of family members show marked divergence in restriction endonuclease sites (Brown and Dover, 1981; Meunier-Rotival *et al.*, 1981). *Eco*RI-LINES probably occur on many mouse chromosomes. Homologous families that largely retain the internal *Eco*RI segment have been detected in other species of *Mus,* in *Apodemus* genomes, and in rats and Chinese hamsters (Heller and Arnheim, 1980; Brown and Dover, 1981). The relative abundance of different variants of the *Eco*RI-LINE family, however, differs in most of these species.

Preliminary data on other possible families of LINES have been published. One, estimated to be repeated about 4×10^4 times, comprises part of the nontranscribed spacer region in the array of mouse genes for ribosomal RNA and also occurs flanking genes for the constant region of mouse heavy chain immunoglobulin (Arnheim *et al.*, 1980). This family is not homologous to the 1.3-kbp *Eco*RI segment of the *Eco*RI-LINE family. Two relatively abundant nonsatellite fragment sets 1.5 and 1.7 kbp long, respectively, and estimated to comprise together about 0.2% of the genome are produced by *Eco*RII cleavage (Manuelidis, 1980); neither one hybridizes with the other. Preliminary evidence indicates the presence of an abundant LINE family in the Kangaroo rat (Liu and Lark, 1981).

3. *Are LINES functional?*

The discovery of LINE families in mammals is recent and there is very little information available regarding function. Adams *et al.* (1980) found no transcripts homologous to the human *Kpn*-LINE family in bone marrow cells and Manuelidis (1981b) also reports negative preliminary experiments. There is no information available regarding the possibility that LINES are mobile in mammalian genomes.

E. Amplification and Dispersion

SINE family members are dispersed between genes, in introns, and within satellite DNA sequences. The SINES are also highly conserved within genomes and between relatively closely related mammalian groups such as members of the genus *Mus* or old world monkeys and man. This conservation is in marked contrast to the very different, noncross-hybridizing (under stringent conditions) satellites of, for example, various old world monkeys and man. On the other hand, when the SINE families of primates and rodents are compared, major differences are apparent even though the repeat units have some striking homologies. There are markedly different multiplicities of different SINE families in different genomes. Further, the B1-SINES of the mouse are, for example, about 130bp in length while the *Alu*-SINES of primates are more than twice that length. At some time after the rodent and primate evolutionary lines separated, one or the other or both lines acquired distinct families of SINES. The mechanism by which the acquisition occurred necessarily involved amplification into many dispersed positions. Providing experimental data relevant to an understanding of that mechanism(s) is a major challenge for future research. The conceptual problems have already been discussed (Scherer and Davis, 1980; Pan *et al.*, 1981; Dover, 1981; Baltimore, 1981; Grimaldi *et al.*, 1981). Unequal crossing-over, which can account for the amplification of satellite DNA (Section III,J), is less satisfactory with regard to amplification and dispersal of SINES.

The process called gene conversion may provide a better model. Gene conversion is a nonreciprocal recombination; a DNA sequence is duplicated at a (partly) homologous distinct genomic site without being lost from the original "donor" site. Both intrachromosomal (Scherer and Davis, 1980) and interchromosomal (Jackson and Fink, 1981; Klein and Petes, 1981) gene conversion have been demonstrated in yeast. Multiple gene conversions might account for the mass change from one SINE family to another partly homologous SINE family as well as the maintenance of homogeneity in an established SINE family. The different relative reiteration frequencies of SINE families might reflect differential rates of gene conversion in different species, but might also reflect more fundamental distinctions between genomes. However, by itself gene conversion does not easily explain the initial or any continuing dispersion of SINE family members into the genome. Dispersion could be explained if SINES prove to be mobile (or transposable) elements, but the mechanism of transposition would have to be such that the donor sequence remains in its original site in addition to being duplicated at a new one (as in gene conversion). The transposable elements of prokaryotes appear to act in just such a way (reviewed by Calos and Miller, 1980).

Similar considerations apply to the amplification and dispersion of LINES. Also, it will be important to understand the processes that maintain LINES in smaller numbers of copies than SINES. Recent data suggest that LINES may be associated with major components of mouse DNA (nonsatellite) that are separable by density gradient centrifugation (Soriano *et al.*, 1981; Meunier-Rotival *et al.*, 1981). Perhaps, unlike SINES, LINES are restricted to particular genomic regions. Second, the extensive polymorphism of LINES within a given species is interesting. Does it reflect a fluid, continually changing population or is it a fixed distribution? And what, if any, are the functional consequences of the polymorphism?

V. Concluding Remarks

The precise insights into the molecular structure and organization of highly repeated sequences that are afforded by modern techniques are rewarding but frustrating. This is especially so in mammals compared, for example, to *Drosophila* (Spradling and Rubin, 1981), because of the difficulties in wedding genetics or cytogenetics with molecular analysis. New experimental approaches to the question of function are essential. One such approach, the search for proteins that specifically bind repeated sequences, has been urged by Brutlag. A protein that interacts specifically with one satellite from *Drosophila melanogaster* has been described (Hsieh and Brutlag, 1979b), but unfortunately this approach has not yet been applied to other organisms. Recent progress in elucidat-

ing the structure of mammalian kinetochores provides interesting possibilities for studying proteins that might bind to satellites (Ris and Witt, 1981; Brenner *et al.*, 1981). Another avenue yet to be fully exploited involves analysis of the effects of repeated sequences on replication and transcription of small, constructed genomes. Genomes for use in mammalian cells have been designed (Hamer, 1980; Mulligan and Berg, 1980) and recombinant DNA techniques permit the introduction of repeated sequences. It is to be hoped that an emphasis on function will develop in the near future.

ACKNOWLEDGMENTS

I am grateful to the following investigators who provided me with preprints: Molly Fitzgerald-Hayes, John Carbon, Joe Maio, Giorgio Bernardi, Prescott Deininger, K. G. Lark, Carl Schmid, Laura Manuelidis, Hans Zachau, A. Plucienniczak, John Rogers, Rolf Streeck. My colleagues Antonella Maresca, Giovanna Grimaldi, Theresa N. H. Lee, and Ronald E. Thayer, kindly and critically reviewed the manuscript.

REFERENCES

Adams, J. W., Kaufman, R. E., Kretschmer, P. J., Harrison, M., and Nienhuis, A. W. (1980). *Nucleic Acids Res.* **8,** 6113–6128.

Alt, F. W., Kellems, R. D., Bertino, J. R., and Schimke, R. T. (1978). *J. Biol. Chem.* **253,** 1357–1370.

Alwine, J. C., Kemp, D. J., Parker, B. A., Reiser, J., Renart, J., Stark, G. R., and Wahl, G. M. (1979). *Methods Enzymol.* **68,** 220–242.

Appels, R., and Peacock, W. J. (1978). *Int. Rev. Cytol. Suppl.* **8,** 69–126.

Arnheim, N., Seperack, P., Banerji, J., Lang, R. B., Miesfeld, R., and Marcu, K. B. (1980). *Cell* **22,** 179–185.

Baltimore, D. (1981). *Cell* **24,** 592–594.

Baralle, F. E., Shoulders, C. C., Goodbourn, S., Jeffreys, A., and Proudfoot, N. (1980). *Nucleic Acids Res.* **8,** 4393–4404.

Barrie, P. A., Jeffreys, A. J., and Scott, A. F. (1981). *J. Mol. Biol.* **149,** 319–336.

Beauchamp, R. S., Mitchell, A. R., Buckland, R. A., and Bostock, C. J. (1979). *Chromosoma* **71,** 153–166.

Bedbrook, J. R., and Gerlach, W. L. (1980). *In* "Genetic Engineering" (J. K. Setlow and A. Hollaender, eds.), Vol. 2, pp. 1–19. Plenum, New York.

Bell, G. I., Pictet, R., and Rutter, W. J. (1980). *Nucleic Acids Res.* **8,** 4091–4109.

Bonner, J., Garrard, W. T., Gottesfeld, J., Holmes, D. S., Sevall, J. S., and Wilkes, M. (1973). *Cold Spring Harbor Symp. Quant. Biol.* **38,** 303–310.

Bostock, C. J. (1980). *Trends Biochem. Sci.* **5,** 117–119.

Bostock, C. J., and Clark, E. M. (1980). *Cell* **19,** 709–715.

Bostock, C. J., Gosden, J. R., and Mitchell, A. R. (1978). *Nature (London)* **272,** 324–328.

Botchan, M. R. (1974). *Nature (London)* **251,** 288–292.

Botchan, M., Topp, W., and Sambrook, J. (1978). *Cold Spring Harbor Symp. Quant. Biol.* **43,** 709–719.

Brenner, S., Pepper, D., Berns, M. W., Tan, E., and Brinkley, B. R. (1981). *J. Cell Biol.* **91,** 95–102.
Britten, R. J., and Davidson, E. H. (1969). *Science* **165,** 349.
Britten, R. J., and Kohne, D. E. (1968). *Science* **161,** 529–540.
Britten, R. J., Graham, D. E., and Neufeld, B. R. (1974). *Methods Enzymol.* **29,** 363–405.
Brown, S. D. M., and Dover, G. (1979). *Nucleic Acids Res.* **6,** 2423–2434.
Brown, S. D. M., and Dover, G. (1980a). *Nucleic Acids Res.* **8,** 781–792.
Brown, S. D. M., and Dover, G. (1980b). *Nature (London)* **285,** 47–49.
Brown, S. D. M., and Dover, G. (1981). *J. Mol. Biol.* **150,** 441–466.
Brutlag, D. L. (1980). *Annu. Rev. Genet.* **14,** 121–144.
Brutlag, D. L., Fry, K., Nelson, T., and Hung, P. (1977). *Cell* **10,** 509–519.
Calos, M. P., and Miller, J. H. (1980). *Cell* **20,** 579–595.
Carlson, M., and Brutlag, D. L. (1977). *Cell* **11,** 371–381.
Cavalier-Smith, T. (1980). *Nature (London)* **285,** 617–618.
Cheng, S.-M., and Schildkraut, C. L. (1980). *Nucleic Acids Res.* **8,** 4075–4090.
Christie, N. T., and Skinner, D. M. (1980). *Proc. Natl. Acad. Sci. U.S.A.* **77,** 2786–2790.
Clarke, L., and Carbon, J. (1980a). *Nature (London)* **287,** 504–509.
Clarke, L., and Carbon, J. (1980b). *Proc. Natl. Acad. Sci. U.S.A.* **77,** 2173–2177.
Clarke, L., Fitzgerald-Hayes, M., Buhler, J.-M., and Carbon, J. (1981). Stadler Symposium, Vol. 13. Univ. of Missouri, Columbia, Missouri, in press.
Coggins, L. W., Grindlay, G. J., Vass, J. K., Slater, A. A., Montague, P., Stinson, M. A., and Paul, J. (1980). *Nucleic Acids Res.* **8,** 3319–3333.
Cooke, H. J., and Hindley, J. (1979). *Nucleic Acids Res.* **6,** 3177–3197.
Cooke, H. J., and McKay, R. D. G. (1978). *Cell* **13,** 453–460.
Corneo, G., Ginelli, E., and Polli, E. (1968). *J. Mol. Biol.* **33,** 331–335.
Corneo, G., Ginelli, E., and Polli, E. (1970). *J. Mol. Biol.* **48,** 319–327.
Corneo, G., Ginelli, E., and Polli, E. (1971). *Biochim. Biophys. Acta* **247,** 528–534.
Corneo, G., Nelli, L. C., Meazza, D., and Ginelli, E. (1980). *Biochim. Biophys. Acta* **607,** 438–444.
Dalla-Favera, R., Gelmann, E. P., Gallo, R. C., and Wong-Staal, F. (1981). *Nature (London)* **292,** 31–35.
Davidson, E. H., and Britten, R. J. (1979). *Science* **204,** 1052–1059.
Davidson, E. H., Hough, B. R., Amenson, C. S., and Britten, R. J. (1973). *J. Mol. Biol.* **77,** 1–23.
Davies, K. E., Young, B. D., Elles, R. G., Hill, M. E., and Williamson, R. (1981). *Nature (London)* **293,** 374–376.
Dawid, I. B., Long, E. O., DiNocera, P. P., and Pardue, M. L. (1981). *Cell* **25,** 399–408.
Deininger, P., and Schmid, C. W. (1976a). *J. Mol. Biol.* **106,** 773–790.
Deininger, P. L., and Schmid, C. W. (1976b). *Science* **194,** 846–848.
Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T., and Schmid, C. W. (1981). *J. Mol. Biol.* **151,** 17–33.
Dennis, E. S., Dunsmuir, P., and Peacock, W. J. (1980). *Chromosoma* **79,** 179–198.
Dhruva, B. R., Shenk, T., and Subramanian, K. N. (1980). *Proc. Natl. Acad. Sci. U.S.A.* **77,** 4514–4518.
Diaz, M. O., Barsacchi-Pilone, G., Mahon, K. A., and Gall, J. G. (1981). *Cell* **24,** 649–659.
Dolnick, B. J., Berenson, R. J., Bertino, J. R., Kaufman, R. J., Nunberg, J. K., and Schimke, R. T. (1979). *J. Cell Biol.* **83,** 395–402.
Donehower, L., and Gillespie, D. (1979). *J. Mol. Biol.* **134,** 805–834.
Donehower, L., Furlong, C., Gillespie, D., and Kurnit, D. (1980). *Proc. Natl. Acad. Sci. U.S.A.* **77,** 2129–2133.
Doolittle, W. F., and Sapienza, C. (1980). *Nature (London)* **284,** 601–603.

Dover, G. (1978). *Nature (London)* **272,** 123–124.
Dover, G. (1980). *Nature (London)* **285,** 618–620.
Dover, G. (1981). *In* "Mechanisms of Speciation" (C. Barigozzi, G. Montalenti, and M. J. D. White, eds.), Rome, in press.
Dover, G., and Doolittle, W. F. (1980). *Nature (London)* **288,** 646–647.
Duhamel-Maestracci, N., Simard, R., Harbers, K., and Spencer, J. H. (1979). *Chromosoma* **75,** 63–74.
Duncan, C., Biro, P. A., Choudary, P. U., Elder, J. T., Wang, R. R. C., Forget, B. G., de Riet, J. K., and Weissman, S. M. (1979). *Proc. Natl. Acad. Sci. U.S.A.* **76,** 5095–5099.
Duncan, C. M., Jagadeeswaran, P., Wang, R. R. C., and Weissman, S. (1981). *Gene* **13,** 185–196.
Dunsmuir, P. (1976). *Chromosoma* **56,** 111–125.
Elder, J. T., Pan, J., Duncan, C. H., and Weissman, S. M. (1981). *Nucleic Acids Res.* **9,** 1171–1189.
Federoff, N., Wellauer, P. K., and Wall, R. (1977). *Cell* **10,** 597–610.
Fittler, F. (1977). *Eur. J. Biochem.* **74,** 343–352.
Fitzgerald-Hayes, M., Buhler, J.-M., Cooper, T. G., and Carbon, J. (1982). *Mol. Cell. Biol.* **2,** 82–87.
Flavell, R. (1980). *Annu. Rev. Plant Physiol.* **31,** 569–596.
Fowlkes, D. M., and Shenk, T. (1980). *Cell* **22,** 405–413.
Fritsch, E. F., Lawn, R. M., and Maniatis, T. (1980). *Cell* **19,** 959–972.
Fritsch, E. F., Shen, C. K. J., Lawn, R. M., and Maniatis, T. (1981). *Cold Spring Harbor Symp. Quant. Biol.* **45,** 761–775.
Fry, K., and Salser, W. (1977). *Cell* **12,** 1069–1084.
Fuke, M., and Busch, H. (1979). *FEBS Lett.* **99,** 136–140.
Fuke, M., Davis, F. M., and Busch, H. (1979). *FEBS Lett.* **102,** 46–50.
Gall, J. G., Stephenson, E. C., Erba, H. P., Diaz, M. O., and Barsacchi-Pilone, G. (1981). *Chromosoma* **84,** 159–171.
Gaillard, G., Doly, J., Cortadas, J., and Bernardi, G. (1981). *Nucleic Acids Res.* **9,** 6069–6082.
Georgiev, G. P., Ilyin, Y. V., Chmeliauskaite, V. G., Ryskov, A. P., Kramerov, D. A., Skryabin, K. G., Krayev, A. S., Lukanidin, E. M., and Grigoryan, M. S. (1981). *Cold Spring Harbor Symp. Quant. Biol.* **45,** 641–654.
Gillespie, D. (1977). *Science* **196,** 889–891.
Gillespie, D., Pequignot, E., and Strayer, D. (1980). *Gene* **12,** 103–111.
Graf, H. (1979). *Hoppe-Seyler Z. Physiol. Chem.* **360,** 1029.
Graf, H., Fittler, F., and Zachau, H. G. (1979). *Gene* **5,** 93–110.
Grimaldi, G., and Singer, M. F. (1982). *Proc. Natl. Acad. Sci. U.S.A.,* in press.
Grimaldi, G., Queen, C., and Singer, M. F. (1981). *Nucleic Acids Res.* **9,** 5553–5568.
Haigwood, N. L., John, C. L., Hutchison, C. A., III, and Edgell, M. H. (1981). *Nucleic Acids Res.* **9,** 1133–1150.
Hamer, D. (1980). *In* "Genetic Engineering" (J. K. Setlow and A. Hollaender, eds.), Vol. 2. Plenum, New York.
Harada, F., and Kato, N. (1980). *Nucleic Acids Res.* **8,** 1273–1285.
Haynes, S. R., Toomey, T. P., Leinwand, L., and Jelinek, W. R. (1981). *Mol. Cell. Biol.* **1,** 573–583.
Heller, R., and Arnheim, N. (1980). *Nucleic Acids Res.* **8,** 5031–5042.
Hilliker, A. J., Appels, R., and Schalet, A. (1980). *Cell* **21,** 607–619.
Hoeijmakers-van Dommelen, H. A. M., Grosveld, G. C., de Boer, E., and Flavell, R. A. (1980). *J. Mol. Biol.* **140,** 531–547.
Hörz, W., and Altenburger, W. (1981). *Nucleic Acids Res.* **9,** 683–696.
Hörz, W., and Zachau, H. G. (1977). *Eur. J. Biochem.* **73,** 383–392.

Hörz, W., Hess, I., and Zachau, H. G. (1974). *Eur. J. Biochem.* **45,** 501–510.
Houck, C. M., and Schmid, C. W. (1981). *J. Mol. Evol.* **17,** 148–155.
Houck, C. M., Rinehart, F. P., and Schmid, C. W. (1979). *J. Mol. Biol.* **132,** 289–306.
Hsiao, C.-L., and Carbon, J. (1981). *Proc. Natl. Acad. Sci. U.S.A.* **78,** 3760–3764.
Hsieh, T., and Brutlag, D. (1979a). *J. Mol. Biol.* **135,** 465–481.
Hsieh, T., and Brutlag, D. (1979b). *Proc. Natl. Acad. Sci. U.S.A.* **76,** 726–730.
Hubbell, H. R., Robberson, D. L., and Hsu, T. C. (1979). *Nucleic Acids Res.* **7,** 2439–2456.
Igo-Kemenes, T., Hörz, W., and Zachau, H. G. (1982). *Annu. Rev. Biochem.* **51,** in press.
Jackson, J. A., and Fink, G. R. (1981). *Nature (London)* **292,** 306–311.
Jahn, C. L., Hutchison, C. A., III, Phillips, S. J., Weaver, S., Haigwood, N. L., Voliva, C. F., and Edgell, M. H. (1980). *Cell* **21,** 159–168.
Jelinek, W. R. (1978). *Proc. Natl. Acad. Sci. U.S.A.* **75,** 2679–2683.
Jelinek, W. R., and Leinwand, L. (1978). *Cell* **15,** 205–214.
Jelinek, W. R., Evans, R., Wilson, M., Salditt-Georgieff, M., and Darnell, J. E. (1978). *Biochemistry* **17,** 2776–2783.
Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L., and Schmid, C. W. (1980). *Proc. Natl. Acad. Sci. U.S.A.* **77,** 1398–1402.
John, B., and Miklos, G. L. G. (1979). *Int. Rev. Cytol.* **58,** 1–114.
Kaput, J., and Sneider, T. W. (1979). *Nucleic Acids Res.* **7,** 2303–2322.
Kaufman, R. E., Kretschmer, P. J., Adams, J. W., Coon, H. C., Anderson, W. F., and Nienhuis, A. W. (1980). *Proc. Natl. Acad. Sci. U.S.A.* **77,** 4229–4233.
Kit, S. (1961). *J. Mol. Biol.* **3,** 711–716.
Klein, H. L., and Petes, T. D. (1981). *Nature (London)* **289,** 144–148.
Kopecka, H., Macaya, G., Cortadas, J., Thiery, J. P., and Bernardi, G. (1978). *Eur. J. Biochem.* **84,** 189–195.
Kornberg, R. (1981). *Nature (London)* **292,** 579–580.
Kramerov, D. A. Grigoryan, A. A., Ryskov, A. P., and Georgiev, G. P. (1979). *Nucleic Acids Res.* **6,** 679–713.
Krayev, A. S., Kremerov, D. A., Skryabin, K. G., Ryskov, A. P., Bayev, A. A., and Georgiev, G. P. (1980). *Nucleic Acids Res.* **8,** 1201–1215.
Kunkel, L. M., Smith, K. D., Boyer, S. H., Borgoankar, D. S., Wachtel, S. S., Miller, O. J., Breg, W. R., Jones, H. W., and Rary, J. M. (1977). *Proc. Natl. Acad. Sci. U.S.A.* **74,** 1245–1249.
Kunkel, L. M., Smith, K. D., and Boyer, S. H. (1979). *Biochemistry* **18,** 3343–3353.
Kurnit, D. M. (1979). *Hum. Genet.* **47,** 169–186.
Kurnit, D. M., and Maio, J. J. (1973). *Chromosoma* **42,** 23–36.
Kurnit, D. M., and Maio, J. J. (1974). *Chromosoma* **45,** 387–400.
Lapeyre, J. N., and Becker, F. F. (1980). *Biochim. Biophys. Acta* **607,** 23–35.
Lapeyre, J.-N., Beattie, W. G., Dugaiczyk, A., Vizard, D., and Becker, F. F. (1980). *Gene* **10,** 339–346.
Lauer, J., Shen, C-K. J., and Maniatis, T. (1980). *Cell* **20,** 119–130.
Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G., and Maniatis, T. (1978). *Cell* **15,** 1157–1174.
Lerner, M. R., and Steitz, J. A. (1981). *Cell* **25,** 298–300.
Lipchitz, L., and Axel, R. (1976). *Cell* **9,** 355–364.
Liu, L.-S., and Lark, K. G. (1981). *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **40,** 1648.
Long, E. O., and Dawid, I. B. (1980). *Annu. Rev. Biochem.* **49,** 727–764.
Macaya, G., Thiery, J.-P., and Bernardi, G. (1977). *In* "Molecular Structure of Human Chromosomes" (J. J. Yunis, ed.), pp. 35–58. Academic Press, New York.
Macaya, G., Cortadas, J., and Bernardi, G. (1978). *Eur. J. Biochem.* **84,** 179–188.

McCutchan, T., Hsu, H., Thayer, R. E., and Singer, M. F. (1982). *J. Mol. Biol.*, in press.
Maio, J. J. (1971). *J. Mol. Biol.* **56,** 579–595.
Maio, J. J., Brown, F. L., and Musich, P. R. (1977). *J. Mol. Biol.* **117,** 637–655.
Maio, J. J., Brown, F. L., and Musich, P. R. (1981a). *Chromosoma* **83,** 103–125.
Maio, J. J., Brown, F. L., McKenna, W. G., and Musich, P. R. (1981b). *Chromosoma* **83,** 127–144.
Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K., and Efstratiadis, A. (1978). *Cell* **15,** 687–701.
Manning, J., Schmid, C.; and Davidson, N. (1975). *Cell* **4,** 141–155.
Manuelidis, L. (1976). *Nucleic Acids Res.* **3,** 3063–3075.
Manuelidis, L. (1978) *Chromosoma* **66,** 1–21.
Manuelidis, L. (1980). *Nucleic Acids Res.* **8,** 3247–3258.
Manuelidis, L. (1981a). *FEBS Lett.* **129,** 25–28.
Manuelidis, L. (1981b). *In* "Genome Evolution and Phenotypic Variation" (G. A. Dover and R. B. Flavell, eds.), Academic Press, in press.
Manuelidis, L., and Wu, J. C. (1978). *Nature (London)* **276,** 92–94.
Marx, K. A., Allen, J. R., and Hearst, J. E. (1976). *Biochim. Biophys. Acta* **425,** 129–147.
Marx, K. A., Purdom, I. F., and Jones, K. W. (1979). *Chromosoma* **73,** 153–161.
Matthews, H. R., Pearson, M. D., and MacLean, N. (1980). *Biochim. Biophys. Acta* **606,** 228–235.
Maxam, A., and Gilbert, W. (1980). *Methods Enzymol.* **65,** 499–560.
Mazrimas, J. A., and Hatch, F. T. (1977). *Nucleic Acids Res.* **4,** 3215–3227.
Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F., and Bernardi, G. (1982). *Proc. Natl. Acad. Sci. U.S.A.* **79,** 355–359.
Miklos, G. L. G., and John, B. (1979). *Am. J. Hum. Genet.* **31,** 264–280.
Miklos, G. L. G., Willcocks, D. A., and Baverstock, P. R. (1980). *Chromosoma (Berlin)* **76,** 339–363.
Mitchell, A. R., Beauchamp, R. S., and Bostock, C. J. (1979). *J. Mol. Biol.* **135,** 127–149.
Mitchell, A. R., Gosden, J. R., and Ryder, O. A. (1981). *Nucleic Acids Res.* **9,** 3235–3249.
Moore, G. P., Constantini, F. D., Posakony, J. W., Davidson, E. H., and Britten, R. J. (1980). *Science* **208,** 1046–1048.
Morrow, J. F. (1979). *Methods Enzymol.* **68,** 3–24.
Mulligan, R., and Berg, P. (1980). *Science* **209,** 1423–1427.
Musich, P. R., Brown, F. L., and Maio, J. J. (1980). *Chromosoma* **80,** 331–348.
Musti, A. M., Sobieski, D., Chen, B. B., and Eden, F. (1981). *Biochemistry* **20,** 2989–2999.
Orgel, L. E., and Crick, F. H. C. (1980). *Nature (London)* **284,** 604–607.
Orgel, L. E., Crick, F. H. C., and Sapienza, C. (1980). *Nature (London)* **288,** 645–646.
Page, G. S., Smith, S., and Goodman, H. M. (1981). *Nucleic Acids Res.* **9,** 2087–2104.
Pan, J., Elder, J. T., Duncan, C., and Weissman, S. M. (1981). *Nucleic Acids Res.* **9,** 1151–1170.
Pardue, M. L., and Gall, J. G. (1970). *Science* **168,** 1356–1358.
Pathak, S., and Würster-Hill, D. H. (1977). *Cytogenet. Cell Genet.* **18,** 245–254.
Pech, M., Streeck, R. E., and Zachau, H. G. (1979a). *Cell* **18,** 883–893.
Pech, M., Igo-Kemenes, T., and Zachau, H. G. (1979b). *Nucleic Acids Res.* **7,** 417–432.
Petes, T. D. (1980). *Cell* **19,** 765–774.
Philippsen, P., Streeck, R. E., and Zachau, H. G. (1974). *Eur. J. Biochem.* **45,** 479–488.
Posakony, J. W., Scheller, R. H., Anderson, D. M., Britten, R. J., and Davidson, E. H. (1981). *J. Mol. Biol.* **149,** 41–67.
Pöschl, E., and Streeck, R. E. (1980). *J. Mol. Biol.* **143,** 147–153.
Queen, C., and Korn, L. (1980). *Methods Enzymol.* **65,** 595–609.
Rinehart, F. P., Ritch, T. G., Deininger, P. L., and Schmid, C. W. (1981). *Biochemistry* **20,** 3003–3010.
Ris, H., and Witt, P. L. (1981). *Chromosoma* **82,** 153–170.

Roberts, R. J. (1980). *Methods Enzymol.* **65,** 1–15.
Roeder, G. S., and Fink, G. R. (1980). *Cell* **21,** 239–249.
Rogers, J. C. (1981), *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **40,** 1649.
Roizes, G. P. (1974). *Nucleic Acids Res.* **1,** 1099–1120.
Roizes, G. P. (1976). *Nucleic Acids Res.* **3,** 2677–2696.
Roizes, G. P., Pages, M., and Lecou, C. (1980). *Nucleic Acids Res.* **8,** 3779–3792.
Rosenberg, H., Singer, M. F., and Rosenberg, M. (1978). *Science* **200,** 394–402.
Rubin, C. M., Deininger, P. L., Houck, C. M., and Schmid, C. W. (1980a). *J. Mol. Biol.* **136,** 151–167.
Rubin, C. M., Houck, C. M., Deininger, P. L., Friedmann, T., and Schmid, C. W. (1980b). *Nature (London)* **284,** 372–374.
Sadler, J. R., Tecklenburg, M., and Betz, J. L. (1980). *Gene* **8,** 279–300.
Sakano, H., Maki, R., Kurosawa, Y., Roeder, W., and Tonegawa, S. (1980). *Nature (London)* **286,** 676–683.
Salser, W., Bowen, S., Browne, D., El Adli, F., Federoff, N., Fry, K., Heindell, H., Paddock, G., Poon, R., Wallace, B., and Whitcomb, P. (1976). *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **35,** 23–35.
Scherer, S., and Davis, R. W. (1980). *Science* **209,** 1380–1384.
Schimke, R. T., Brown, P. C., Kaufman, R. J., McGrogan, M., and Slate, D. L. (1980). *Cold Spring Harbor Symp. Quant. Biol.* **45,** 785–797.
Schmeckpeper, B. J., Willard, H. F., and Smith, K. D. (1981). *Nucleic Acids Res.* **9,** 1853–1872.
Schmid, C. W., and Deininger, P. L. (1975). *Cell* **6,** 345–358.
Sealy, L., Hartley, J., Donelson, J., Chalkley, R., Hutchison, N., and Hamkalo, B. (1981). *J. Mol. Biol.* **145,** 291–318.
Segal, S., Garner, M., Singer, M. F., and Rosenberg, M. (1976). *Cell* **9,** 247–257.
Sen, S., and Sharma, J. (1980). *Chromosoma* **81,** 393–402.
Shen, C.-K. J., and Maniatis, T. (1980). *Cell* **19,** 379–391.
Shih, C., Padhy, L. C., Murray, M. J., and Weinberg, R. A. (1981). *Nature (London)* **290,** 261–264.
Singer, D. S. (1979). *J. Biol. Chem.* **254,** 5506–5514.
Singer, D. S., and Donehower, L. (1979). *J. Mol. Biol.* **134,** 835–842.
Smith, A. J. H. (1980). *Methods Enzymol.* **65,** 560–580.
Smith, G. P. (1976). *Science* **191,** 528–535.
Smith, T. F. (1980). *Nature (London)* **285,** 620.
Soriano, P., Macaya, G., and Bernardi, G. (1981). *Eur. J. Biochem.* **115,** 235–239.
Southern, E. M. (1970). *Nature (London)* **227,** 794–798.
Southern, E. M. (1975a). *J. Mol. Biol.* **94,** 51–69.
Southern, E. M. (1975b). *J. Mol. Biol.* **98,** 503–517.
Southern, E. M. (1979). *Methods Enzymol.* **68,** 152–176.
Spradling, A. C., and Rubin, G. M. (1981). *Annu. Rev. Genet.* **15,** 219–264.
Stambrook, P. J. (1981). *Biochemistry* **20,** 4393–4398.
Streeck, R. E. (1981). *Science* **213,** 443–445.
Streeck, R. E., and Zachau, H. G. (1978). *Eur. J. Biochem.* **89,** 267–279.
Streeck, R. E., Pech, M., and Zachau, H. G. (1979). *In* "FEBS Special Meeting on Enzymes" (P. Mildner, ed.). Pergamon, Oxford.
Sturm, K. S., and Taylor, J. H. (1981). *Nucleic Acids Res.* **9,** 4537–4546.
Szostak, J. W., and Wu, R. (1980). *Nature (London)* **284,** 426–430.
Szybalski, W. (1968). *Methods Enzymol.* **12B,** 330–360.
Tartof, K. D. (1975). *Annu. Rev. Genet.* **9,** 355–385.
Tashima, M., Calabretta, B., Tovelli, G., Scofield, M., Maizel, A., and Saunders, G. F. (1981). *Proc. Natl. Acad. Sci. U.S.A.* **78,** 1508–1512.
Temin, H. M. (1980). *Cell* **21,** 599–600.

Thayer, R. E., McCutchan, T., and Singer, M. F. (1981). *Nucleic Acids Res.* **9,** 169–181.
Varley, J. M., MacGregor, H. C., and Erba, H. P. (1980a). *Nature (London)* **283,** 686–688.
Varley, J. M., MacGregor, H. C., Nardi, I., Andrews, C., and Erba, H. P. (1980b). *Chromosoma* **80,** 289–307.
Wahl, G. M., Padgett, R. A., and Stark, G. R. (1979). *J. Biol. Chem.* **254,** 8679–8689.
Weiner, A. M. (1980). *Cell* **22,** 209–218.
Wensink, P. C., Tabata, S., and Pachl, C. (1979). *Cell* **18,** 1231–1246.
Wu, J. C., and Manuelidis, L. (1980). *J. Mol. Biol.* **142,** 363–386.
Zachau, H. G., and Igo-Kemenes, T. (1981). *Cell* **24,** 597–598.
Zieve, G. W. (1981). *Cell* **25,** 296–297.

Note Added in Proof

The mouse interspersed repeated family described by Arnheim *et al.* (1980) is probably a SINE family, not a LINE family [Miesfeld, R., Krystal, M., and Arnheim, N. (1981). *Nucleic Acids Res.* **9,** 5931–5947].